MACROMOLECULE MASS SPECTROMETRY: CITATION MINING OF USER DOCUMENTS

By

Dr. Ronald N. Kostoff, Office of Naval Research
Arlington, VA, 22217, USA
Phone: 703-696-4198; Fax: 703-696-4274
Internet: kostofr@onr.navy.mil

Dr. Clifford D. Bedford, Office of Naval Research
Arlington, VA, 22217, USA

Dr. J. Antonio del Río and Héctor D. Cortes
Centro de Investigación en Energía, UNAM
Temixco, Mor. México

Dr. George Karypis
University of Minnesota
Minneapolis, MN 55455, USA

(*THE VIEWS IN THIS REPORT ARE SOLELY THOSE OF THE AUTHORS, AND DO NOT NECESSARILY REPRESENT THE VIEWS OF THE U.S. DEPARTMENT OF THE NAVY OR ANY OF ITS COMPONENTS, THE UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO, OR THE UNIVERSITY OF MINNESOTA.*)

KEYWORDS: Bibliometrics; Citation Mining; Text Mining; Mass Spectrometry; MALDI; Electrospray; Research Impact; Citation Analysis; Metrics; Soft Laser Desorption; Nobel Prize; Time-of-Flight; FTMS; Quadrupole Ion Trap.

1) ABSTRACT

Identifying research users, applications, and impact is important for research performers, managers, evaluators, and sponsors. It is important to know whether the audience reached is the audience desired. It is useful to understand the technical characteristics of the other research/ development/ applications impacted by the originating research, and to understand other characteristics (names, organizations, countries) of the users impacted by the research. Because of the many indirect

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **14 NOV 2003**	2. REPORT TYPE	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **MACROMOLECULE MASS SPECTROMETRY CITATION MINING OF USER DOCUMENTS**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Ronald Kostoff; Clifford Bedford ; Jesus Del Rio; Hector Cortes; George Karypis;**	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of Naval Research,800 N. Quincy St.,Arlington,VA,22217,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Identifying research users, applications, and impact is important for research performers, managers, evaluators, and sponsors. It is important to know whether the audience reached is the audience desired. It is useful to understand the technical characteristics of the other research/ development/ applications impacted by the originating research, and to understand other characteristics (names, organizations, countries) of the users impacted by the research. Because of the many indirect pathways through which fundamental research can impact applications, identifying the user audience and the research impacts can be very complex and time consuming. This report identified the literature pathways through which two highly-cited papers of 2002 Chemistry Nobel Laureates Fenn and Tanaka impacted other research, technology development, and applications. It also identified the technical and infrastructure characteristics of the user population. Citation Mining, an integration of citation bibliometrics and text mining, was applied to the >1600 first generation Science Citation Index (SCI) citing papers to FennŸs 1989 Science paper on Electrospray Ionization for Mass Spectrometry, and to the >400 first generation SCI citing papers to TanakaŸs 1988 Rapid Communications in Mass Spectrometry paper on Laser Ionization Time-of-Flight Mass Spectrometry. Bibliometrics was performed on the citing papers to profile the user characteristics. Text mining was performed on the citing papers to identify the technical areas impacted by the research, the relationships among these technical areas, and relationships among the technical areas and the infrastructure (authors, journals, organizations).

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**89**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

pathways through which fundamental research can impact applications, identifying the user audience and the research impacts can be very complex and time consuming.

This report identified the literature pathways through which two highly-cited papers of 2002 Chemistry Nobel Laureates Fenn and Tanaka impacted other research, technology development, and applications. It also identified the technical and infrastructure characteristics of the user population.

Citation Mining, an integration of citation bibliometrics and text mining, was applied to the >1600 first generation Science Citation Index (SCI) citing papers to Fenn's 1989 Science paper on Electrospray Ionization for Mass Spectrometry, and to the >400 first generation SCI citing papers to Tanaka's 1988 Rapid Communications in Mass Spectrometry paper on Laser Ionization Time-of-Flight Mass Spectrometry. Bibliometrics was performed on the citing papers to profile the user characteristics.
Text mining was performed on the citing papers to identify the technical areas impacted by the research, the relationships among these technical areas, and relationships among the technical areas and the infrastructure (authors, journals, organizations).

2) BACKGROUND

This report applies modern information technology techniques to a subset of the macromolecular mass spectrometry literature. The Background section describes the growth of the Electrospray Ionization and Laser Desorption Mass Spectrometry literatures, and relates the growth of these literatures to the original papers by Nobel co-recipients John B. Fenn and Koichi Tanaka, and to the papers of other principal contributors as well. The Background section then proceeds to describe the information technology approaches used in this analysis (text mining, bibliometrics, citation mining).

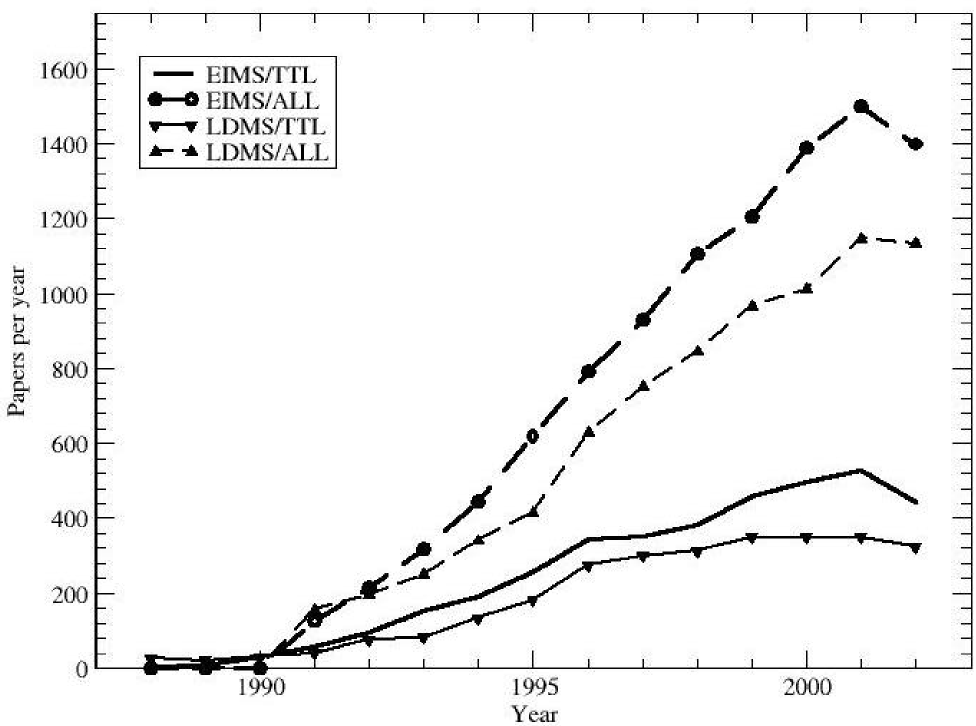
a. **Growth of the Macromolecular Mass Spectrometry Literature**

The 2002 Nobel Prize in Chemistry was shared by John B. Fenn, Koichi Tanaka, and Kurt Wuthrich for their work in developing methods to enable the identification and structural analysis of biological macromolecules. In particular, Fenn and Tanaka focused on soft desorption ionization methods. Fenn concentrated on electrospray ionization [1-7], and Tanaka concentrated on soft laser desorption [8-10].

The impact of these researchers on their respective disciplines can be viewed from a literature perspective. Figure 1 shows the growth in the SCI Electrospray Ionization Mass Spectrometry (EIMS) literature (retrieved by the query Electrospray AND (Mass OR Ion* OR Spectrometry)), and the growth in the Laser Desorption Mass Spectrometry (LDMS) literature (retrieved by the query Laser AND Desorption AND (Ion* OR Mass Spectrometry) from 1988 to mid-2002. The dashed curves are based on papers retrieved by a query applied to all text fields (Title, Abstract, Keywords), while the solid curves are based on a query applied to the Title field only. Before 1991, Abstracts were not available for SCI papers.

FIGURE 1 – GROWTH IN ELECTROSPRAY AND LASER DESORPTION LITERATURES

(Papers per Year vs Time)



In the years that EIMS growth accelerated initially (1988-1990), essentially all the papers retrieved from the database cited one or more of Fenn's papers dating from 1984 [1-7]. From the 'bottom-up' perspective, references [1-7] received a total of 151 citations between 1984 and 1990, of which 143 were from external groups. The top twenty of these 143 citing papers received over 150 citations apiece, with an aggregate second-generation citation total (for these top twenty alone) of 5400 citations.

In the years that LDMS growth accelerated initially (1990-1992), 145 papers were retrieved from the title search only. The top fifty cited papers of the 145 retrieved ranged in citations from 983 to 33. Tanaka's 1988 paper [8] was referenced in fifteen, one or more of R. C. Beavis' papers [e.g., 11-13] were referenced in 37, and one or more of M. Karas' papers [e.g., 14-15] were referenced in 38 of these top fifty cited papers. Many of these Karas papers were published jointly with F. Hillenkamp. Reference [14] in particular has received over 1450 citations to date. From the 'bottom-up' perspective, reference [8] received a total of 69 citations between 1988 and 1992, of which all were from external groups. The top fourteen of these 69 citing papers received over 100 citations apiece, with an aggregate second-generation citation total (for these top fourteen alone) of 3140 citations.

References [1-8] have been cited highly. In particular, references [1-7] have received ~590, 210, 670, 210, 370, 1630, 890 citations respectively, by November 2002, and reference [8] has received 410 citations. The citing community can be viewed as a sub-set of the total user community. Identifying the characteristics of the citing community would provide one perspective on the diversity of *impact* that these papers have had or, more accurately, on the diversity of *citings* that these papers have had.

b. Text Mining

Science and technology (S&T) text mining [16-19] is a computational linguistics-based process for extracting useful information from large volumes of technical text. It identifies pervasive technical themes in large databases from frequently occurring technical phrases. It also identifies relationships among these themes by grouping (clustering) these phrases (or their parent documents) on the basis of similarity. Text mining can be used for:

- Enhancing information retrieval and increasing awareness of the global technical literature [20-22]
- Potential discovery and innovation based on merging common linkages between very disparate literatures [23-26]

- Uncovering unexpected asymmetries from the technical literature [27-28]
- Estimating global levels of effort in S&T sub-disciplines [29-31]
- Helping authors potentially increase their citation statistics by improving access to their published papers, and thereby potentially helping journals to increase their Impact Factors [32]
- Tracking myriad research impacts across time and applications areas [33-34].

A typical text mining study of the published literature develops a query for comprehensive information retrieval, processes the database using computational linguistics and bibliometrics, and integrates the processed information.

c. Bibliometrics

Evaluative bibliometrics [35-37] uses counts of publications, patents, citations and other potentially informative items to develop science and technology performance indicators. Its validity is based on the premises that 1) counts of patents and papers provide valid indicators of R&D activity in the subject areas of those patents or papers, 2) the number of times those patents or papers are cited in subsequent patents or papers provides valid indicators of the impact or importance of the cited patents and papers, and 3) the citations from papers to papers, from patents to patents and from patents to papers provide indicators of intellectual linkages between the organizations which are producing the patents and papers, and knowledge linkage between their subject areas [38]. Evaluative bibliometrics can be used to:

- Identify the infrastructure (authors, journals, institutions) of a technical domain,
- Identify experts for innovation-enhancing technical workshops and review panels,
- Develop site visitation strategies for assessment of prolific organizations globally,
- Identify impacts (literature citations) of individuals, research units, organizations, and countries

d. Citation Mining

Citation Mining [34, 39] is a technique developed for the purpose of characterizing the aggregate citing papers of a research unit. A research unit can consist of one paper, selected papers from an author, or selected papers from a group or technical discipline. In Citation Mining, text mining and bibliometrics analyses are performed on the aggregate citing papers. The bibliometrics component yields the infrastructure information (e.g., prolific authors, journals, institutions, countries, most cited authors, papers, journals, etc), and the computational linguistics component yields the

pervasive technical thrusts and the relationships among the thrusts. A temporal component documents the dissemination of information to the research and user community as a function of time.

The Science Citation Index (SCI) is a database that links papers (P1) in journals indexed by the SCI to other SCI papers (P2) that cite the original papers P1, and contains references (P3) in the original papers P1 as well. While the SCI accesses many of the premier research journals, it does not access all technical journals published. In the present study, the SCI is used to identify the citing papers to Fenn's and Tanaka's original papers. Thus, all the citing papers in the technical literature will not be identified, only those in journals accessed by the SCI.

This report describes the application of Citation Mining to the subset of the most highly cited papers of Fenn [6] and Tanaka [8] referenced above, using the SCI as the source for citing papers. It was desired to examine papers that were cited highly, preferably with multi-discipline readership journals where possible, to obtain the broadest potential areas for application. Because the SCI did not use Abstracts until 1991, and because Abstract analysis is a key feature of Citation Mining, it was desired to examine papers published relatively close to 1991. Because temporal dissemination and impacts of the initial cited papers is also a key feature of Citation Mining, it was desired to limit the analysis to one paper from each researcher, in order to have a sharp starting point in time. Therefore, references [6] and [8] were selected as the seeds for the Citation Mining process.

Section 3 presents the Results, divided into a bibliometrics sub-section and a computational linguistics sub-section. Section 4 presents the Summary, section 5 presents the Conclusions, and section 6 contains the References.

3) RESULTS

The results from the publications bibliometric analyses are presented in section 3.1, followed by the results from the citations bibliometrics analysis in section 3.2. Results from the computational linguistics analyses are shown in section 3.3. The SCI bibliometric fields incorporated into the database included, for each paper, the author, journal, institution, Keywords, and references. Reference [40] contains an abridged version of the complete study results.

3.1 Publication Statistics on Authors, Journals, Organizations, Countries

The first group of metrics presented is counts of papers published by different entities. These metrics can be viewed as output and productivity measures. They are not direct measures of research quality, although there is some threshold quality level inferred, since these papers are published in the (typically) high caliber journals accessed by the SCI.

There were 1628 papers that cited Fenn's 1989 paper, and 410 papers that cited Tanaka's 1988 paper. Because the SCI did not start to publish Abstracts until 1991, and because not all citing papers have Abstracts, only 1433 of the Fenn citing papers in the SCI database contain Abstracts, and only 344 of the Tanaka citing papers contain Abstracts. The bibliometrics analyses are performed on the total number of citing papers, whereas the computational linguistics are performed on those papers with Abstracts.

3.1.1. Author Frequency Results

The 1628 Fenn citing papers contain 3602 different authors, and 6263 author listings, resulting in 3.8 author listings per paper. The 410 Tanaka citing papers contain 973 different authors and 1462 different author listings, resulting in 3.57 author listings per paper. The occurrence of each author's name on a paper is defined as an author listing. The number of author listings per paper is relatively high in either case, and seems to follow a trend set by earlier text mining studies. In four previous chemistry-related text mining studies, this ratio averaged over 3.5, while in three previous fluid mechanics-related text mining studies, this ratio averaged under 2.5. A high value of this ratio tends to indicate large teams characteristic of large experimental efforts, while a low value of this ratio tends to indicate small teams characteristic of individual theoretical or computational modeling efforts. The most prolific authors of the Fenn citing papers are listed in Table 1A, and the most prolific authors of the Tanaka citing papers are listed in Table 1B.

TABLE 1A – MOST PROLIFIC AUTHORS – FENN CITING PAPERS
(present institution listed)

AUTHOR	INSTITUTION	COUNTRY	# PAPERS
SMITH—RD	PACIFIC NW NATL LAB	USA	48
MCLUCKEY—SA	PURDUE UNIV	USA	43
MCLAFFERTY—FW	CORNELL UNIV	USA	42
LOO—JA	PFIZER GLOBAL R&D	USA	37
CLEMMER—DE	INDIANA UNIV	USA	34
COLTON—R	LA TROBE UNIV	AUSTRALIA	34
MANN—M	UNIV SO DENMARK	DENMARK	29

MUDDIMAN—DC	VCU	USA	26
ROEPSTORFF—P	ODENSE UNIV	DENMARK	26
TRAEGER—JC	LA TROBE UNIV	AUSTRALIA	26
WILLIAMS—ER	UNIV CAL BERKELEY	USA	22
HENION—JD	CORNELL UNIV	USA	20
MARSHALL—AG	FLORIDA STATE UNIV	USA	19
ARAKAWA—R	KANSAI UNIV	JAPAN	18
COUNTERMAN—AE	INDIANA UNIV	USA	18
STEPHENSON—JL	RES TRIANGLE INST	USA	18
VANBERKEL—GJ	OAK RIDGE NATL LAB	USA	18
CHAIT—BT	ROCKEFELLER UNIV	USA	17
LITTLE—DP	SEQUENOM, INC	USA	15
EDMONDS—CG	PACIFIC NW NATL LAB	USA	14
JOHNSON—RS	IMMUNEX R&D CORP	USA	14
SENKO—MW	FLORIDA STATE UNIV	USA	14

TABLE 1B – MOST PROLIFIC AUTHORS – TANAKA CITING PAPERS

AUTHOR	INSTITUTION	COUNTRY	# PAPERS
ZENOBI—R	SWISS FED INST TECH	SWITZERLAND	18
HILLENKAMP—F	UNIV MUNSTER	GERMANY	12
KARAS—M	UNIV FRANKFURT	GERMANY	12
COTTER—RJ	JHU	USA	11
GROTEMEYER—J	UNIV KIEL	GERMANY	9
KNOCHENMUSS—R	SWISS FED INST TECH	SWITZERLAND	9
WILKINS—CL	UNIV ARKANSAS	USA	9
DERRICK—PJ	UNIV WARWICK	UK	8
HERCULES—DM	VANDERBILT UNIV	USA	8
AMSTER—IJ	UNIV GEORGIA	USA	7
RUSSELL—DH	TEXAS A&M UNIV	USA	7
BAHR—U	JW GOETHE UNIV	GERMANY	6
BURLINGAME—AL	UNIV CAL SAN FRANCISCO	USA	6
CASTORO—JA	UNIV CAL RIVERSIDE	USA	6
DEAK—G	DEBRECEN UNIV MED	HUNGARY	6
FENSELAU—C	UNIV MARYLAND	USA	6
KEKI—S	LAJOS KOSSUTH UNIV	HUNGARY	6
KUHN—G	FED INST MAT RES & TEST	GERMANY	6
PERERA—IK	UNIV HULL	UK	6
SCHLAG—EW	TECH INST MUNCHEN	GERMANY	6
SUNDQVIST—BUR	UNIV UPPSALA	SWEDEN	6
WEIDNER—S	FED INST MAT RES & TEST	GERMANY	6
ZSUGA—M	DEBRECEN UNIV MED	HUNGARY	6

These regional distributions are very different. For the Fenn citing papers, seventeen of the 22 most prolific authors are from the USA, two are from Australia, two are from Denmark, and one is from Japan. Fifteen are from universities, three are from research institutes, and four are from industry.

For the Tanaka citing papers, eight of the 23 most prolific authors are from the USA, and the remainder are from Europe, mainly central Europe. Twenty are from universities, and three are from research institutes. No authors are common to the two lists of prolific citing authors. Why are there no prolific citing authors from Japan, and why are there no prolific citing authors from industry, for Tanaka's research? This is surprising, since Tanaka is both from Japan and industry.

Two notes of caution. First, the institutions listed are typically the most recent at which the author can be found. Since many researchers have cycled through a number of institutions globally over the course of their careers, the author numbers may not compare exactly with the institution or country numbers shown later. Second, separate listing of authors does not mean that the papers are separate. For example, most, if not all, of the papers by Hillenkamp and Karas in Table 1B are co-authored.

3.1.2 Journal frequency results

There were 317 different journals represented in the Fenn citing papers, with an average of 5.14 papers per journal. There were 112 different journals represented in the Tanaka citing papers, with an average of 3.67 papers per journal. These ratios are about half the values as the previous chemistry text mining studies, but on the same order as the previous fluid mechanics text mining studies. The previous text mining studies were thematic (i.e., all the papers had the common themes of the search query), while the present aggregation of citing papers is not thematic in the same sense. Given the thematic focus of many technical journals, it is reasonable that the citing papers would be distributed over a wider group of journals, with a wider aggregate thematic base. The journals containing the most Fenn citing papers are listed in Table 2A, and the journals containing the most Tanaka citing papers are listed in Table 2B.

TABLE 2A – JOURNALS CONTAINING MOST FENN CITING PAPERS

JOURNAL	# PAPERS
ANALYTICAL CHEMISTRY	193
JOURNAL OF THE AMERICAN SOCIETY FOR MASS SPECTROMETRY	139
RAPID COMMUNICATIONS IN MASS SPECTROMETRY	132
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY	72
JOURNAL OF MASS SPECTROMETRY	68
ANALYTICAL BIOCHEMISTRY	37
INTERNATIONAL JOURNAL OF MASS SPECTROMETRY	33

JOURNAL OF CHROMATOGRAPHY A	29
INTERNATIONAL JOURNAL OF MASS SPECTROMETRY AND ION PROCESSES	26
BIOCHEMISTRY	25
JOURNAL OF BIOLOGICAL CHEMISTRY	23
ELECTROPHORESIS	23
INORGANICA CHIMICA ACTA	21
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	20
PROTEIN SCIENCE	19
JOURNAL OF AEROSOL SCIENCE	19
BIOLOGICAL MASS SPECTROMETRY	19
ANALYTICA CHIMICA ACTA	18
MASS SPECTROMETRY REVIEWS	17
EUROPEAN JOURNAL OF BIOCHEMISTRY	17

TABLE 2B – JOURNALS CONTAINING MOST TANAKA CITING PAPERS

JOURNAL	# PAPERS
RAPID COMMUNICATIONS IN MASS SPECTROMETRY	70
ANALYTICAL CHEMISTRY	56
JOURNAL OF THE AMERICAN SOCIETY FOR MASS SPECTROMETRY	34
INTERNATIONAL JOURNAL OF MASS SPECTROMETRY AND ION PROCESSES	20
JOURNAL OF MASS SPECTROMETRY	16
MACROMOLECULES	14
ORGANIC MASS SPECTROMETRY	13
INTERNATIONAL JOURNAL OF MASS SPECTROMETRY	11
JOURNAL OF CHROMATOGRAPHY A	7
FRESENIUS JOURNAL OF ANALYTICAL CHEMISTRY	6
ANALYTICA CHIMICA ACTA	6
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY	5
BIOLOGICAL MASS SPECTROMETRY	5
EUROPEAN MASS SPECTROMETRY	5
JOURNAL OF BIOLOGICAL CHEMISTRY	5
MASS SPECTROMETRY REVIEWS	4
REVIEW OF SCIENTIFIC INSTRUMENTS	4
JOURNAL OF PHYSICAL CHEMISTRY B	4

In both cases, the most prolific journals focus on mass spectrometry, chemistry, and biology. Three journals stand out as the first tier for containing the most citing papers: ANALYTICAL CHEMISTRY, JOURNAL OF THE AMERICAN SOCIETY FOR MASS SPECTROMETRY, RAPID COMMUNICATIONS IN MASS SPECTROMETRY. Twelve journals are in common between the two lists. The Fenn citing journals not in common tend to focus on biology/ biochemistry (ANALYTICAL BIOCHEMISTRY, BIOCHEMISTRY, PROTEIN SCIENCE, EUROPEAN JOURNAL OF BIOCHEMISTRY), while the Tanaka citing journals not in common tend to focus on the

technique/ instrumentation (REVIEW OF SCIENTIFIC INSTRUMENTS, ORGANIC MASS SPECTROMETRY, EUROPEAN MASS SPECTROMETRY). This observation supports the later document clustering finding of the greater emphasis on bio-molecules in the Fenn citing papers relative to the Tanaka citing papers.

3.1.3 Institution frequency results

A similar process was used to develop a frequency count of institutional address appearances. It should be noted that many different organizational components may be included under the single organizational heading (e.g., Harvard Univ could include the Chemistry Department, Biology Department, Physics Department, etc.). Identifying the higher level institutions is instrumental for these DT studies. Once they have been identified through bibliometric analysis, subsequent measures may be taken (if desired) to identify particular departments within an institution.

There were 801 different institutions represented in the Fenn citing papers, with an average of 2.03 papers per institution. There were 315 different institutions represented in the Tanaka citing papers, with an average of 1.3 papers per institution. The institutions producing the most Fenn citing papers are listed in Table 3A, and the institutions producing the most Tanaka citing papers are listed in Table 3B.

TABLE 3A – INSTITUTIONS PRODUCING MOST FENN CITING PAPERS

INSTITUTION	COUNTRY	# PAPERS
CORNELL UNIV	USA	66
OAK RIDGE NATL LAB	USA	52
BATTELLE MEM INST	USA	47
VIRGINIA COMMONWEALTH UNIV	USA	41
YALE UNIV	USA	38
INDIANA UNIV	USA	38
UNIV WASHINGTON	USA	36
LA TROBE UNIV	AUSTRALIA	35
ODENSE UNIV	DENMARK	33
OSAKA UNIV	JAPAN	29
NATL RES COUNCIL CANADA	CANADA	26
UNIV ALBERTA	CANADA	25
PURDUE UNIV	USA	25
UNIV CALIF SAN FRANCISCO	USA	25
UNIV CALIF BERKELEY	USA	22
FLORIDA STATE UNIV	USA	22
UNIV MICHIGAN	USA	18
ROCKEFELLER UNIV	USA	17
NYU	USA	17
CALTECH	USA	17

TABLE 3B – INSTITUTIONS PRODUCING MOST TANAKA CITING PAPERS

INSTITUTION	COUNTRY	# PAPERS
SWISS FED INST TECH	SWITZERLAND	18
UNIV MUNSTER	GERMANY	14
JOHNS HOPKINS UNIV	USA	12
UNIV GEORGIA	USA	11
TECH UNIV MUNICH	GERMANY	9
UNIV CALIF RIVERSIDE	USA	9
UNIV WARWICK	UK	9
UNIV PITTSBURGH	USA	7
UNIV CALIF SAN FRANCISCO	USA	6
UNIV UPPSALA	SWEDEN	6
UNIV VIENNA	AUSTRIA	6
INDIANA UNIV	USA	6
UNIV ILLINOIS	USA	6
CNR	ITALY	6
LOUISIANA STATE UNIV	USA	5
ROHM & HAAS CO	USA	5
ARIZONA STATE UNIV	USA	5
TEXAS A&M UNIV	USA	5
ROCKEFELLER UNIV	USA	5
OSAKA UNIV	JAPAN	5

Of the twenty institutions producing the most Fenn citing papers, seventeen are from North America, one from Europe, and two from the Far East. Seventeen are universities, and three are research institutes. Of the twenty institutions producing the most Tanaka citing papers, twelve are from the USA, seven are from Europe, and one is from Japan. Eighteen are universities, one is a research institute, and one is from industry. Four institutions are in common between the two lists: UNIV CAL SAN FRANCISCO, INDIANA UNIV, ROCKEFELLER UNIV, OSAKA UNIV.

3.1.4 Country frequency results

There are 51 different countries listed in the Fenn citing papers, and 36 different countries listed in the Tanaka citing papers. The countries producing the most Fenn citing papers are listed in Table 4A, and the countries producing the most Tanaka citing papers are listed in Table 4B. The dominance of a handful of countries is clearly evident.

TABLE 4A – COUNTRIES PRODUCING THE MOST FENN CITING PAPERS

COUNTRY	# PAPERS
USA	917
CANADA	119
GERMANY	115
JAPAN	102
ENGLAND	83
FRANCE	80
AUSTRALIA	69
DENMARK	42
NETHERLANDS	36
SWEDEN	35
SWITZERLAND	35
PEOPLES R CHINA	28
ITALY	26
BELGIUM	22
SPAIN	15
RUSSIA	12
SCOTLAND	12
HUNGARY	11
NEW ZEALAND	10
TAIWAN	8

TABLE 4B – COUNTRIES PRODUCING THE MOST TANAKA CITING PAPERS

COUNTRY	# PAPERS
USA	193
GERMANY	48
ENGLAND	33
JAPAN	31
CANADA	23
SWITZERLAND	23
NETHERLANDS	12
FRANCE	11
SWEDEN	10
HUNGARY	8
ITALY	8
AUSTRALIA	6
AUSTRIA	6
SCOTLAND	6
BELGIUM	5
PEOPLES R CHINA	5
ISRAEL	4
RUSSIA	4

The USA clearly dominates in country output. The next tier is high on both lists (GERMANY, ENGLAND, JAPAN, CANADA), with Switzerland appearing high on the Tanaka citing list. Thus, while Japan is not very visible in terms of prolific citing

authors or institutions, especially with respect to Tanaka's paper, it has reasonable representation in terms of country citations. This implies a diverse group of citing authors in Japan, with the exception of the group at Osaka University.

Figure 1A contains a co-occurrence matrix of the top 15 countries listed in the Fenn citing papers, and Figure 1B contains a co-occurrence matrix of the top 15 countries listed in the Tanaka citing papers.

FIGURE 1A – COUNTRY CO-OCCURRENCE MATRIX FOR FENN CITING PAPERS

COUNTRY	AUSTRAL	BELGIUM	CANADA	DENMARK	ENGLAND	FRANCE	GERMANY	ITALY	JAPAN	HOLLAND	CHINA	SPAIN	SWEDEN	SWITZER	USA
AUSTRALIA	69	0	1	0	3	1	2	0	0	0	0	0	0	0	1
BELGIUM	0	22	0	0	0	3	1	0	0	1	0	1	0	0	1
CANADA	1	0	119	1	4	8	1	0	0	0	1	0	0	0	20
DENMARK	0	0	1	42	0	3	4	0	1	1	0	0	1	0	4
ENGLAND	3	0	4	0	83	4	3	1	1	4	0	0	0	3	15
FRANCE	1	3	8	3	4	80	2	1	1	1	0	0	1	3	11
GERMANY	2	1	1	4	3	2	115	0	4	1	0	0	1	4	15
ITALY	0	0	0	0	1	1	0	26	0	0	0	0	1	1	2
JAPAN	0	0	0	1	1	1	4	0	102	0	1	0	0	0	16
HOLLAND	0	1	0	1	4	1	1	0	0	36	0	1	1	1	0
CHINA	0	0	1	0	0	0	0	0	1	0	28	0	2	0	0
SPAIN	0	1	0	0	0	0	0	0	0	1	0	15	0	0	2
SWEDEN	0	0	0	1	0	1	1	1	0	1	2	0	35	0	5
SWITZERLAND	0	0	0	0	3	3	4	1	0	1	0	0	0	35	5
USA	1	1	20	4	15	11	15	2	16	0	0	2	5	5	917

FIGURE 1B – COUNTRY CO-OCCURRENCE MATRIX FOR TANAKA CITING PAPERS

COUNTRY	AUSTRAL	AUSTRIA	BELGIUM	CANADA	ENGLAND	FRANCE	GERMANY	HUNGARY	ITALY	JAPAN	HOLLAND	CHINA	SCOTLAN	SWEDEN	SWITZER	USA
AUSTRALIA	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AUSTRIA	0	6	0	0	0	1	1	0	0	0	0	0	0	0	0	0

BELGIUM	0	0	5	0	0	0	0	1	0	0	0	0	0	0	0	1
CANADA	0	0	0	23	1	0	1	0	0	0	0	0	0	0	0	6
ENGLAND	0	0	0	1	33	0	1	1	0	0	1	0	1	1	2	4
FRANCE	0	1	0	0	0	11	1	0	0	0	0	0	0	0	0	1
GERMANY	0	1	0	1	1	1	48	1	0	0	0	0	0	0	0	7
HUNGARY	0	0	1	0	1	0	1	8	0	0	0	0	0	0	0	1
ITALY	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	1
JAPAN	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	3
HOLLAND	0	0	0	0	1	0	0	0	0	0	12	0	0	0	0	0
CHINA	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	1
SCOTLAND	0	0	0	0	1	0	0	0	0	0	0	0	6	0	2	0
SWEDEN	0	0	0	0	1	0	0	0	0	0	0	0	0	10	0	2
SWITZERLAND	0	0	0	0	2	0	0	0	0	0	0	0	2	0	23	0
USA	0	0	1	6	4	1	7	1	1	3	0	1	0	2	0	193

In terms of absolute numbers of co-authored Fenn-citing papers, the USA major partners are Canada, Japan, Germany, England, and France. Additionally, the USA is the major partner for ten of the countries, the exceptions being Australia, Belgium, Holland, and China.

In terms of absolute numbers of co-authored Tanaka-citing papers, the USA major partners are Germany, Canada, England, and Japan. Additionally, the USA is the major partner for nine of the countries, the exceptions being Australia, Austria, Holland, Scotland, and Switzerland.

3.2 Citation Statistics on Authors, Papers, and Journals

The second group of metrics presented is counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics [36], *much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers* [41, 42].

The citations in all the retrieved SCI papers were aggregated, the authors, specific papers, years, journals, and countries cited most frequently were identified, and were presented in order of decreasing frequency. A small percentage of any of these categories received large numbers of citations.

3.2.1 Author citation frequency results

The most highly cited authors in the Fenn citing papers are listed in Table 5A, and the most highly cited authors in the Tanaka citing papers are listed in Table 5B. These

represent the authors who are highly co-cited with Fenn and Tanaka, respectively. Only the first authors of the cited papers in the Fenn citing papers are listed.

TABLE 5A – MOST CITED AUTHORS IN FENN CITING PAPERS
(cited by other papers in this database only)

AUTHOR	INSTITUTION	COUNTRY	# CITES
FENN JB	VCU	USA	1982
SMITH RD	PACIFIC NW NATL LAB	USA	1134
LOO JA	PFIZER GLOBAL R&D	USA	875
KARAS M	UNIV FRANKFURT	GERMANY	600
MCLUCKEY SA	PURDUE UNIV	USA	541
MANN M	UNIV SO DENMARK	DENMARK	450
BIEMANN K	MIT	USA	343
CHOWDHURY SK	SANOFI WINTHROP INC	USA	302
COVEY TR	SCIEX LTD	CANADA	297
KATTA V	AMGEN INC	USA	287
YAMASHITA M	TOKAI UNIV	JAPAN	285
HUNT DF	UNIV VIRGINIA	USA	279
VANBERKEL GJ	OAK RIDGE NATL LAB	USA	266
COLTON R	LA TROBE UNIV	AUSTRALIA	258
MARSHALL AG	FLORIDA STATE UNIV	USA	252
MCLAFFERTY FW	CORNELL UNIV	USA	239
HILLENKAMP F	UNIV MUNSTER	GERMANY	235
GANEM B	CORNELL UNIV	USA	217
BRUINS AP	UNIV GRONINGEN	NETHERLANDS	211
WILM M	EUROPEAN MOL BIOL LAB	GERMANY	203
BEAVIS RC	NYU	USA	202

TABLE 5B – MOST CITED AUTHORS IN TANAKA CITING PAPERS
(cited by other papers in this database only)

AUTHOR	INSTITUTION	COUNTRY	# CITES
KARAS M	UNIV FRANKFURT	GERMANY	659
BEAVIS RC	NYU	USA	422
TANAKA K	SHIMADZU CORP	JAPAN	410
HILLENKAMP F	UNIV MUNSTER	GERMANY	242
SPENGLER B	UNIV GIESSEN	GERMANY	201
DANIS PO	ROHM AND HAAS CO	USA	143
MONTAUDO G	UNIV PISA	ITALY	134
COTTER RJ	JHU	USA	114
VERTES A	GWU	USA	111
FENN JB	VCU	USA	102
NELSON RW	INTRINS BIOPROBES INC	USA	97
BARBER M	UMIST	UK	94
OVERBERG A	UNIV MUNSTER	GERMANY	89
SMITH RD	PACIFIC NW NATL LAB	USA	82
BOESL U	TECH UNIV MUNICH	GERMANY	75

JUHASZ P	PERCEPT BIOSYS	USA	70
STRUPAT K	UNIV MUNSTER	GERMANY	69
CHAIT BT	ROCKEFELLER UNIV	USA	69
GROTEMEYER J	UNIV KIEL	GERMANY	64
LI L	UNIV ALBERTA	CANADA	61
BENNINGHOVEN A	UNIV MUNSTER	GERMANY	61

In the Fenn citing papers, Fenn is cited almost twice as much as the next ranked author. This is due to the citation of Fenn's other related papers between 1984 and 1989 [1-5, 7-8], in addition to the citation of the Science article [6]. The next tier, RD Smith and JA Loo, was a very prolific and highly cited group working on different mass spectrometry techniques, including electrospray ionization [e.g., 43-45].

In the Tanaka citing papers, Tanaka actually ranks third in number of first-author citations. M. Karas of Frankfurt ranks first (along with F. Hillenkamp of Muenster, who co-authored many of these papers with Karas). This is due to three factors. First, in 1985, Karas, in conjunction with Hillenkamp, showed that a "strongly absorbing matrix at a fixed laser wavelength" could be used to vaporize small molecules without chemical degradation [46]. Second, in 1988, Karas and Hillenkamp reported a MALDI approach applied to proteins [47] shortly after Tanaka's paper was published. Thus, the papers that cite Tanaka's paper also tend to cite the groundwork papers of Karas/ Hillenkamp as well as their large molecule mass determination papers. Third, Karas and Hillenkamp were in the top tier of Tanaka citing authors, as well as prolific in their own right relative to Tanaka, and had more opportunity to cite their own foundational work in the papers in which they also cited Tanaka [e.g., 48]. Additionally, due to a series of highly-cited papers by R.C. Beavis (along with his co-author B. Chait) in the early 1990s on laser desorption mass spectrometry [e.g., 11-13], many of the papers that cite Tanaka tend to multiply cite Beavis/ Chait. This large co-citation of Karas/ Hillenkamp and Beavis/ Chait with Tanaka was mentioned in the Background. It was shown that, of the top fifty cited laser desorption mass spectrometry papers produced in the early high growth years, Tanaka's paper was referenced in fifteen, while the Beavis/ Chait papers were referenced in 37 and the Karas/ Hillenkamp papers were referenced in 38.

There are five names in common between the two lists (FENN, SMITH, KARAS, BEAVIS, HILLENKAMP). All these five have made broad contributions to mass spectrometry.

Of the 21 most cited authors in the Fenn citing papers, fourteen are from universities, three are from research institutions, and four are from industry. Of the 21 most cited

authors in the Tanaka citing papers, sixteen are from universities, one is from a research institute, and four are from industry. This relatively high fraction (~20%) of cited papers from industry suggests relatively applied citing papers. The validity of this assumption is confirmed in the sections on temporal citing patterns and document clustering.

Finally, while Central Europe plays a modest role in the reference source for the Fenn list, it continues to play a much stronger role for the Tanaka list.

The citation data for authors and journals represents citations generated only by the specific records extracted from the SCI database for this study. It does not represent all the citations received by the references in those records; these references in the database records could have been cited additionally by papers in other technical disciplines.

3.2.2 Document citation frequency results

The most highly cited documents in the Fenn citing papers are listed in Table 6A, and the most highly cited documents in the Tanaka citing papers are listed in Table 6B.

TABLE 6A – MOST CITED DOCUMENTS IN FENN CITING PAPERS
(total citations listed in SCI)

AUTHOR	YEAR	JOURNAL	VOLUME	TOT CITES
FENN JB	1989	SCIENCE	V246,P64	1628
ELECTROSPRAY IONIZATION FOR MASS-SPECTROMETRY OF LARGE BIOMOLECULES				
SMITH RD	1990	ANAL CHEM	V62,P882	854
BIOCHEMICAL MASS-SPECTROMETRY - ELECTROSPRAY IONIZATION				
KARAS M	1988	ANAL CHEM	V60,P2299	1329
LASER DESORPTION IONIZATION OFLARGE PROTEINS				
FENN JB	1990	MASS SPECTROM REVIEW	V9,P37	879
ELECTROSPRAY IONIZATION				
SMITH RD	1991	MASS SPECTROM REVIEW	V10,P359	482
ELECTROSPRAY IONIZATION MASS SPECTROMETRY FOR LARGE POLYPEPTIDES				
COVEY TR	1988	RAPID COMM MASS SPEC	V2,P249	486
PROTEIN MOLECULAR WEIGHTS BY ION SPRAY MASS SPECTROMETRY				
YAMASHITA M	1984	J PHYS CHEM	V88,P4451	576
ELECTROSPRAY ION-SOURCE - FREE-JET THEME				
WHITEHOUSE CM	1985	ANAL CHEM	V57,P675	653
ELECTROSPRAY INTERFACE FOR LIQUID CHROMATOGRAPHS AND MASS SPECTROMETERS				
HILLENKAMP F	1991	ANAL CHEM	V63,PA1193	983
MATRIX-ASSISTED LASER DESORPTION IONIZATION MASS-SPECTROMETRY OF BIOPOLYMERS				
MANN M	1989	ANAL CHEM	V61,P1702	361

MASS-SPECTRA OF MULTIPLY CHARGED IONS				
BRUINS AP	1987	ANAL CHEM	V59,P2642	619
LIQUID CHROMATOGRAPHY/ATMOSPHERIC PRESSURE IONIZATION MASS-SPECTROMETRY				
DOLE M	1968	J CHEM PHYS	V49,P2240	357
MOLECULAR BEAMS OF MACROIONS				
ROEPSTORFF P	1984	BIOMED MASS SPECTROM	V11,P601	1058
COMMON NOMENCLATURE FOR SEQUENCE IONS IN MASS-SPECTRA OF PEPTIDES				
CHOWDHURY SK	1990	J AM CHEM SOC	V112,P9012	230
PROBING CONFORMATIONAL-CHANGES IN PROTEINS BY MASS-SPECTROMETRY				
CHOWDHURY SK	1990	RAPID COMM MASS SPEC	V4,P81	223
ELECTROSPRAY-IONIZATION MASS-SPECTROMETER				
WILM MS	1994	INT J MASS SPECTROM	V136,P167	286
ELECTROSPRAY AND TAYLOR-CONE THEORY, DOLES BEAM OF MACROMOLECULES				
GANEM B	1991	J AM CHEM SOC	V113,P6294	248
DETECTION OF NONCOVALENT RECEPTOR LIGAND COMPLEXES BY MASS-SPECTROMETRY				
HUNT DF	1986	P NATL ACAD SCI USA	V83,P6233	530
PROTEIN SEQUENCING BY TANDEM MASS-SPECTROMETRY				
IRIBARNE JV	1976	J CHEM PHYS	V64,P2287	313
EVAPORATION OF SMALL IONS FROM CHARGED DROPLETS				

TABLE 6B – MOST CITED DOCUMENTS IN TANAKA CITING PAPERS

(total citations listed in SCI)

AUTHOR	YEAR	JOURNAL	VOLUME	TOT CITES
TANAKA K	1988	RAPID COMM MASS SPEC	V2,P151	410
LASER IONIZATION TIME-OF-FLIGHT MASS SPECTROMETRY				
KARAS M	1988	ANAL CHEM	V60,P2299	1329
LASER DESORPTION IONIZATION OFLARGE PROTEINS				
KARAS M	1987	INT J MASS SPECTROM	V78,P53	574
MATRIX-ASSISTED ULTRAVIOLET-LASER DESORPTION OF NONVOLATILE COMPOUNDS				
HILLENKAMP F	1991	ANAL CHEM	V63,PA1193	983
MATRIX-ASSISTED LASER DESORPTION IONIZATION MASS-SPECTROMETRY OF BIOPOLYMERS				
BEAVIS RC	1989	RAPID COMM MASS SPEC	V3,P233	233
ULTRAVIOLET LASER DESORPTION OF PROTEINS				
BEAVIS RC	1990	ANAL CHEM	V62,P1836	276
PROTEIN MOLECULAR MASS USING MATRIX-ASSISTED LASER DESORPTION MASS-SPECTROMETRY				
BEAVIS RC	1989	RAPID COMM MASS SPEC	V3,P432	357
CINNAMIC ACID DERIVATIVES MATRICES FOR UV LASER DESORPTION MASS SPECTROMETRY				
FENN JB	1989	SCIENCE	V246,P64	1628
ELECTROSPRAY IONIZATION FOR MASS-SPECTROMETRY OF LARGE BIOMOLECULES				
BEAVIS RC	1991	CHEM PHYS LETT	V181,P479	217
VELOCITY DISTRIBUTIONS OF INTACT HIGH MASS POLYPEPTIDE MOLECULE IONS PRODUCED BY MATRIX ASSISTED LASER DESORPTION				
BAHR U	1992	ANAL CHEM	V64,P2866	270
MASS-SPECTROMETRY OF SYNTHETIC-POLYMERS BY UV MATRIX-ASSISTED LASER DESORPTION IONIZATION				
STRUPAT K	1991	INT J MASS SPECTROM	V111,P89	263
LASER DESORPTION/ IONIZATION MASS SPECTROMETRY				
SPENGLER B	1990	ANAL CHEM	V62,P793	115

ULTRAVIOLET-LASER DESORPTION IONIZATION MASS-SPECTROMETRY OF LARGE PROTEINS BY PULSED ION EXTRACTION TIME-OF-FLIGHT ANALYSIS

DANIS PO	1992 ORG MASS SPECTROM	V27,P843	158
ANALYSIS OF WATER-SOLUBLE POLYMERS BY MATRIX-ASSISTED LASER DESORPTION TIME-OF-FLIGHT MASS-SPECTROMETRY			
FENN JB	1990 MASS SPECTROM REV	V9,P37	879
ELECTROSPRAY IONIZATION			
OVERBERG A	1990 RAPID COMM MASS SPEC	V4,P293	113
INFRARED MATRIX-ASSISTED LASER DESORPTION/ IONIZATION MASS SPECTROMETRY			
BEAVIS RC	1990 P NATL ACAD SCI USA	V87,P6873	225
ANALYSIS OF PROTEIN MIXTURES BY MASS-SPECTROMETRY			
DANIS PO	1993 ORG MASS SPECTROM	V28,P923	133
SAMPLE PREPARATION FOR THE ANALYSIS OF SYNTHETIC ORGANIC POLYMERS BY MATRIX-ASSISTED LASER-DESORPTION IONIZATION			
BARBER M	1981 J CHEM SOC CHEM COMM	P325	1024
FAST ATOM BOMBARDMENT OF SOLIDS (FAB) - A NEW ION-SOURCE FOR MASS-SPECTROMETRY			
WILEY WC	1955 REV SCI INSTRUM	V26,P1150	1537
TIME-OF-FLIGHT MASS SPECTROMETER WITH IMPROVED RESOLUTION			
CASTRO JA	1992 RAPID COMM MASS SPEC	V6,P239	115
MATRIX-ASSISTED LASER DESORPTION IONIZATION OF HIGH-MASS MOLECULES BY FOURIER-TRANSFORM MASS-SPECTROMETRY			

The theme of each paper is shown in italics on the line after the paper listing. The order of paper listings is by number of citations by other papers in the extracted database analyzed. The total number of citations from the SCI paper listing, a more accurate measure of total impact, is shown in the last column on the right.

For the twenty most cited documents in the Fenn citing papers, Analytical Chemistry contains the most highly cited documents (six). For the twenty most cited documents in the Tanaka citing papers, both Analytical Chemistry and Rapid Communications in Mass Spectrometry each contain five of the most highly cited documents.

All of the journals containing these most highly cited documents are fundamental science journals, and most of the topics have a fundamental science theme. Of the most highly cited documents in the Fenn citing papers, nine are from the 80s, eight are from the 90s, and one each from the 70s and 60s. Of the most highly cited documents in the Tanaka citing papers, twelve are from the 90s, seven are from the eighties, and one is from the 50s. These numbers reflect dynamically evolving disciplines, with many of the seminal works coming from recent times.

From Table 6A, about thirty percent of the papers address the phenomena underlying electrospray (ION SOURCE-FREE JET, ELECTROSPRAY INTERFACE, MULTIPLY-CHARGED IONS, MACROION BEAMS, CHARGED DROPLET ION EVAPORATION),

about twenty five percent address the electrospray technique (ELECTROSPRAY IONIZATION, HYBRID MASS SPECTROMETRY), about thirty percent address applications (LARGE POLYPEPTIDES, PROTEINS, RECEPTOR LIGAND COMPLEXES), and a few address laser desorption. From Table 6B, about fifteen percent of the papers address the laser desorption approach and associated phenomena, about ten percent address the electrospray technique, and the remainder address applications (LARGE PROTEINS, NONVOLATILE COMPOUNDS, BIOPOLYMERS, LARGE BIOMOLECULES, SYNTHETIC POLYMERS), mainly using the MALDI technique. The relatively large numbers of cited papers related to applications are consistent with the observation in the previous section that a relatively substantial number of highly cited authors were from industrial organizations.

3.2.3. Journal citation frequency results

The most highly cited journals in the Fenn citing papers are listed in Table 7A, and the most highly cited journals in the Tanaka citing papers are listed in Table 7B.

TABLE 7A – MOST CITED JOURNALS IN FENN CITING PAPERS
(cited by other papers in this database only)

JOURNAL	# CITES
ANAL CHEM	8699
J AM CHEM SOC	4550
RAPID COMMUN MASS SP	3888
J AM SOC MASS SPECTR	3371
SCIENCE	3006
INT J MASS SPECTROM	2010
J BIOL CHEM	1809
P NATL ACAD SCI USA	1701
BIOCHEMISTRY-US	1305
MASS SPECTROM REV	1231
ANAL BIOCHEM	1141
J MASS SPECTROM	1076
ELECTROPHORESIS	1069
J PHYS CHEM-US	1020
J CHEM PHYS	965
J CHROMATOGR	965
ORG MASS SPECTROM	935
NATURE	888
METHOD ENZYMOL	607
J CHROMATOGR A	550

TABLE 7B – MOST CITED JOURNALS IN TANAKA CITING PAPERS

JOURNAL	# CITES
ANAL CHEM	2895
RAPID COMMUN MASS SP	2471
INT J MASS SPECTROM	1082
J AM SOC MASS SPECTR	652
J AM CHEM SOC	556
ORG MASS SPECTROM	488
J BIOL CHEM	454
SCIENCE	309
BIOMED ENVIRON MASS	293
MACROMOLECULES	285
MASS SPECTROM REV	273
P NATL ACAD SCI USA	257
CHEM PHYS LETT	244
J MASS SPECTROM	225
J CHEM PHYS	213
J PHYS CHEM-US	211
ANAL BIOCHEM	191
BIOL MASS SPECTROM	177
BIOCHEMISTRY-US	152
J CHROMATOGR	134

Sixteen of the top twenty most highly cited journals are in common between the two lists. Those not in common from the list of most cited journals in Fenn citing papers (Table 7A) are: ELECTROPHORESIS, NATURE, METHODS ENZYMOLOGY, JOURNAL OF CHROMATOGRAPHY A. Those not in common from the list of most cited journals in Tanaka citing papers (Table 7B) are: BIOMEDICAL ENVIRONMENTAL MASS, MACROMOLECULES, CHEM PHYS LETTERS, BIOLOGICAL MASS SPECTROMETRY.

The journals containing the most Fenn citing papers (Table 2A) and the most cited journals in the Fenn citing papers (Table 7A) had thirteen journals in common. The journals containing the most Tanaka citing papers (Table 2B) and the most cited journals in the Tanaka citing papers (Table 7B) also had thirteen journals in common.

3.3 Temporal Citing Patterns

In the original citation mining papers [34, 39], two characteristics of the citing papers were evaluated as a function of time. These were: 1) the level of development of the work reported in the citing paper (basic research, applied research, technology development) and 2) the alignment between the technical thrusts of the citing paper and the cited paper (strongly aligned, partially aligned, not aligned). These temporal results provided useful insights to the evolution of the nature of the citing papers as time proceeded, and it was decided to perform a similar analysis for the present paper.

In order to have sufficient data to evaluate these two characteristics credibly, only those citing papers with Abstracts were included in the analysis (1433 citing papers for Fenn, 344 citing papers for Tanaka)

A two character metric was used to quantify the above two characteristics. The first character represented level of development, and ranged from one (most fundamental research) to three (applied technology development). The second character represented degree of alignment, and ranged from one (fully aligned) to three (non-aligned). Table 8 presents the temporal citing results. Table 8A presents the results for the Fenn citing papers (Table 8A-1-normalized), and Table 8B presents the results for the Tanaka citing papers (Table 8B-1-normalized). The first column in each table is the two character metric. The matrix elements Mij represent the number of citing papers with metric i published in year j. For example, in Table 8A, there were 25 citing papers in 2001 that were both fundamental research and fully aligned with the theme of the (Fenn) cited paper.

TABLE 8A – FENN CITING PAPER CHARACTERISTICS VS TIME

METRIC	ALL YEARS	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	451	18	25	30	49	52	68	69	27	36	28	23	24	2
12	328	16	13	12	11	29	44	40	60	44	40	10	8	1
13	121	6	12	9	5	9	10	6	10	21	14	14	5	
21	299	13	21	24	43	42	30	18	21	23	22	19	22	1
22	170	18	38	56	14	10	8	6	7	8	3	1	1	
23	49	16	14	9	3	2	4	1						
31	10	1	2	5	1	1								
32	4		3	1										
33	1		1											
TOTAL->	1433	88	129	146	126	145	164	140	125	132	107	67	60	4

TABLE 8A-1 – FENN CITING PAPER CHARACTERISTICS VS TIME (NORM)

METRIC	ALL YEARS	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	0.31	0.2	0.19	0.21	0.39	0.36	0.41	0.49	0.22	0.27	0.26	0.34	0.4	0.5
12	0.23	0.18	0.1	0.08	0.09	0.2	0.27	0.29	0.48	0.33	0.37	0.15	0.13	0.25
13	0.08	0.07	0.09	0.06	0.04	0.06	0.06	0.04	0.08	0.16	0.13	0.21	0.08	0
21	0.21	0.15	0.16	0.16	0.34	0.29	0.18	0.13	0.17	0.17	0.21	0.28	0.37	0.25
22	0.12	0.2	0.29	0.38	0.11	0.07	0.05	0.04	0.06	0.06	0.03	0.01	0.02	0
23	0.03	0.18	0.11	0.06	0.02	0.01	0.02	0.01	0	0	0	0	0	0
31	0.01	0.01	0.02	0.03	0.01	0.01	0	0	0	0	0	0	0	0
32	0	0	0.02	0.01	0	0	0	0	0	0	0	0	0	0
33	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0

TABLE 8B – TANAKA CITING PAPER CHARACTERISTICS VS TIME

METRIC	ALL YEARS	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	136	9	9	12	6	15	18	15	11	11	14	6	10	0
12	86	9	10	4	10	7	9	9	10	5	6	6	1	
13	50	2	2	5	4	7	10	6	2	5	4	1	2	
21	43	1	1	4	3	4	1	6	5	5	5	3	5	
22	20	1	4	4		3		2	1	2	2	1		
23	6	1	2	1		2								
31	2			2										
32	0													
33	1		1											
TOTAL->	344	23	29	32	23	38	38	38	29	28	31	17	18	0

TABLE 8B-1 – TANAKA CITING PAPER CHARACTERISTICS VS TIME (NORM)

METRIC	ALL YEARS	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	0.4	0.39	0.31	0.38	0.26	0.39	0.47	0.39	0.38	0.39	0.45	0.35	0.56	
12	0.25	0.39	0.34	0.13	0.43	0.18	0.24	0.24	0.34	0.18	0.19	0.35	0.06	
13	0.15	0.09	0.07	0.16	0.17	0.18	0.26	0.16	0.07	0.18	0.13	0.06	0.11	
21	0.13	0.04	0.03	0.13	0.13	0.11	0.03	0.16	0.17	0.18	0.16	0.18	0.28	
22	0.06	0.04	0.14	0.13	0	0.08	0	0.05	0.03	0.07	0.06	0.06	0	
23	0.02	0.04	0.07	0.03	0	0.05	0	0	0	0	0	0	0	
31	0.01	0	0	0.06	0	0	0	0	0	0	0	0	0	
32	0	0	0	0	0	0	0	0	0	0	0	0	0	
33	0	0	0.03	0	0	0	0	0	0	0	0	0	0	

In aggregate, the Tanaka citing papers have a moderately greater concentration in basic research (first metric character of unity) than the Fenn citing papers, 0.80 normalized vs. 0.62 normalized. The Tanaka citing papers have a greater concentration in the most non-aligned category (second metric character of three) than the Fenn citing papers, 0.17 normalized vs. 0.11 normalized. These two findings corroborate the most prolific authors bibliometrics results, which showed almost twenty percent of the most prolific Fenn citing authors were from industry, whereas none of the most prolific Tanaka citing authors were from industry.

The temporal evolution shows that about a decade is required before the applied technology citing papers become evident. It should be stressed that these are the directly citing technology papers, i.e., papers that cited the original Fenn or Tanaka papers. It is possible that indirectly citing technology papers (i.e., papers that did not cite Fenn or Tanaka's original paper, but rather cited other papers that had cited the

Fenn or Tanaka original papers) appeared earlier, but this higher generation bibliometric analysis was beyond the scope of the present study.

One other citation mining study has been performed [34, 39]. Emphasized in that study, and comparable in spirit to the present study, was a detailed analysis of the 1992 Science paper of Jaeger and Nagel on dynamic granular systems [49]. That paper was a very fundamental research paper focused on the basic physics of flowing granular systems. The normalized temporal evolution of the citing papers of that study is shown in Table 9. Relative to the Fenn and Tanaka citing papers, the Jaeger and Nagel citing papers have a substantially higher basic research fraction in aggregate. There was a four year lag time before any applied citing papers emerged. Beyond what the numbers portray, the Jaeger and Nagel citing papers reached a wider variety of more extreme non-aligned categories than the Fenn or Tanaka citing papers (e.g., earthquakes, avalanches, traffic congestion, war games, flow immunosensors, shock waves, nanolubrication, thin film ordering). Chi-tests confirmed the validity of the differences between the Fenn-Tanaka citing papers and the Jaeger and Nagel citing papers, and between the Fenn and Tanaka citing papers as well.

TABLE 9 – JAEGER AND NAGEL CITING PAPER CHARACTERISTICS VS TIME (NORM)

METRIC	ALL YEARS	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	0.78	0.67	0.69	0.75	0.75	0.84	0.77	0.85	0.85	0.75		
12	0.15	0.17	0.21	0.2	0.18	0.08	0.17	0.09	0.07	0		
13	0.04	0	0.02	0	0.05	0.04	0.06	0.06	0.07	0.25		
21	0.01	0	0	0.03	0.02	0	0	0	0	0		
22	0.01	0	0.06	0	0	0.02	0	0	0	0		
23	0.01	0.17	0.02	0.03	0	0	0	0	0	0		
31	0	0	0	0	0	0	0	0	0	0		
32	0	0	0	0	0	0.02	0	0	0	0		
33	0	0	0	0	0	0	0	0	0	0		

3.4. Computational Linguistics (Taxonomy Generation)

Three statistically-based clustering methods, factor matrix, multi-link aggregation, and partitional document clustering, were used to develop taxonomies. They each offered a modestly different perspective on taxonomy category structure. Neither of the three approaches is inherently superior, and all should be viewed as complementary.

3.4.1. Fenn Citing Papers

The words contained in the Fenn citing paper Abstracts were extracted by the Vantage Point software, and sorted by frequency of occurrence. The highest frequency high technical content words were identified by inspection. Very similar words were consolidated (e.g., singulars/ plurals, full spellings/ acronyms, very strong synonyms).

3.4.1.1. Factor Matrix Clustering

A correlation matrix of the 253 resultant consolidated words was generated, and a factor analysis was performed using the WINSTAT statistical package (an Excel add-in). The eigenvalue floor was set equal to unity to insure that each resulting factor provide value-added information, and a factor matrix consisting of 42 factors (columns) and 253 words (rows) resulted.

Each matrix element Mij is known as the factor loading, and is a measure of the contribution of word i to factor j. A factor represents a technical theme, and some combination of the factors represents a taxonomy. There are cases where words have high loadings in multiple factors (e.g., RECOMBINANT has a value of .46 in factor 1, .37 in factor 10, and .21 in factor 37), and they usually (not always) tend to be situated in the factor where they have the highest loading. These high loading multi-factor words do, however, serve as a link among the factors, and cause the factors to overlap.

Overall, the factor matrix required more factors, and had more overlap among factors, than in previous text mining studies. These previous studies focused on papers related to a focused theme, not to a cited paper as in the present study. The citing paper database is more diverse and fragmented, since it incorporates many different types of applications. Its component papers tend to mix both application and technique/ technology development, as opposed to the much stronger focus on technique/ technology development that characterized the previous studies. This added diversity, and the mixing of technology development with applications, translates into a larger number of factors that have numerous overlaps.

The factors in the matrix are ordered by cohesiveness. Factor 1 is larger in extent and more focused than the other factors. As the factor numbers increase, the factors contain less words and their theme becomes more diffuse. Due to space limitations, only the larger factors (1-22) will be analyzed. A description of each factor, and the aggregation of all factors into a taxonomy, follows. The capitalized words in parentheses represent typical high factor loading (>.3) words for the factor described. In some cases, words that have high factor loadings and are physically in close

proximity will be components of multi-word phrases (e.g., AMINO, ACID, SEQUENCES, in first factor).

Factor 1 (AMINO, RESIDUES, ACIDS, PEPTIDES, SEQUENCING, N-TERMINAL, DISULFIDE, C-TERMINAL, PROTEINS, CYSTEINE, TRYPSIN, PROTEOLYSIS, SITES, DIGESTS, PURIFICATION, METHIONINE, RECOMBINANT, ENZYME, MAPPING, CHAINS, BONDS, ISOLATION, TERMINAL, LYSINE, ARGININE, CLEAVAGES, PHOSPHORYLATION, NATIVE, STRUCTURES, SERINE, COLI) – focuses on protein construction, characterization and structure through analysis of digested and recombined amino acid, peptide and polypeptide sequences and residues. General disulfide cleavage, trypsin digestion, and determination of C- and N-terminal groups, afford methods for reconstructive mapping of fragments and segments of proteins and other macromolecules.

Factor 2 (DISSOCIATION, FRAGMENTS, CID, COLLISIONS, PRECURSOR, TANDEM, MS/MS, ENERGY, CLEAVAGES, IONS, SPECTRA, QUADRAPOLE, TRAP, PATTERNS, PROTONATION, ACTIVITY, PARENT, BACKBONE, EXCITATION, SINGLY, MASSES) – focuses on the analysis of species generated from proteins and other macromolecules by collision-induced dissociation in quadrupole ion trap coupled tandem mass spectrometry. Ion mass spectra analysis of the resulting fragments determines the structure (typically) of metabolites, peptides and polypeptides and affords information on the reconstruction of parent molecules or proteins.

Factor 3 (LASER, MALDI-MS, MATRIX-ASSISTED, DESORPTION/IONIZATION, MATRIX, DESORPTION, DECAY, TIME-OF-FLIGHT, DIGESTS, COMBINED, MEMBRANE, APPLICATIONS, PEPTIDES, PROTEINS, SPECTROMETRY) – focuses on matrix-assisted laser desorption ionization time-of-flight mass spectrometry for peptide mass fingerprinting, followed by post source decay analysis for more detailed characterization of amino acid sequences.

Factor 4 (CHROMOTOGRAPHY, LIQUID, HPLC, COLUMN, REVERSED-PHASE, SEPARATION, LC, COUPLING, STANDARDS, LC/MS, EXTRACTS, INJECTION, DIGESTS, CAPILLARY, TRYPSIN, COMBINED, FLOW, ELECTROPHORESIS, MONITORING) – focuses on separation techniques used to purify and separate digested macromolecules, proteins or polypeptides prior to introduction into, and identification by, various mass spectrometry methods. For biochemical materials, two separation techniques are generally used; high pressure liquid chromatography and

gel electrophoresis. This factor reflects the wide use of electrospray ionization for interfacing liquid separation methods with mass spectrometry quickly.

Factor 5 (DROPLETS, DIAMETER, SURFACE, FLOW, RATES, SPRAY, EVAPORATION, ELECTRIC, SIZE, ELECTROSTATIC, CONE, CONCENTRATIONS, LIMITS, CHARGES, LIQUID, CAPILLARY) – focuses on the effect of electrostatic and liquid properties, and flow variables, on the size and charge of droplets ejected in conical sprays from capillaries prior to solvent elimination and injection of the digested substrate material fragments into the mass spectrometer for characterization.

Factor 6 (ELECTROSPRAYED, IONIZATION, MASSES, ESI, SPECTROMETRY, ESI-MS, SPECTRA, DETECTION, IONS, SHIFTS, NONCOVALENT, SPECIES, SOLUTIONS) – focuses on the use of electrospray ionization mass spectrometry for the detection of mass spectra of large molecules in solution.

Factor 7 (CROSS, CONFORMATIONS, STATES, UBIQUITIN, GAS-PHASE, MOBILITY, CYTOCHROME, TUBE, CHARGES, TEMPERATURE, GAS, LYSOZYME, PHASE, EXCHANGE, CHARGE-STATE) – focuses on the interaction between protein digestion (ubiquitin, lysozyme) and aspects of the structure/ conformation determination process using ion mobility mass spectrometry

Factor 8 (METAL, ALKALI, CATIONS, SODIUM, SALTS, SPECIES, LIGANDS, ETHER, ADDUCTS, NEUTRAL, SELECTIVITY, OXYGEN) – focuses on use of negative ion mass spectrometry to study the structure of sodium and other alkali metal salts, with emphasis on adduction of alkali metal compounds with anions of the larger alkali metal ions.

Factor 9 (INTERACTIONS, NONCOVALENT, BINDING, COMPLEXES, DIMER, INHIBITOR, AFFINITY, BONDS, SCREENING, HYDROPHOBIC, STABILITY, LIGANDS, PHASE) – focuses on use of soft ionization mass spectrometry for studying noncovalently bound complexes, including interaction strength. Emphasis is on deduction of the stoichiometry of the binding partners from the molecular weight measurement, and use of the mass spectrometry-based method to assess the affinity of such interactions. Focuses on the non-ionic protein and macromolecule conformational and structural interactions.

Factor 10 (ESCHERICHIA, COLI, EXPRESSION, GENE, PURIFICATION, METHIONINE, NATIVE, RECOMBINANT, MUTANT) – focuses on genetically expressed proteins in Escherichia Coli cells and recombination technology.

Factor 11 (CYCLOTRON, FOURIER, RESONANCE, ISOTOPIC, EXCITATION, RESOLVING, IONS) – focuses on Fourier transform ion cyclotron resonance mass spectrometry, including different ion source configurations and cyclotron resonance excitation, for instruments of higher mass resolution.

Factor 12 (PRESSURE, ATMOSPHERIC, VACUUM, INTERFACE, ANALYZER, ELECTRIC, FOCUSING, CHROMOTOGRAPHY/ MASS) – focuses on the collision-activated dissociation at the vacuum/ atmospheric pressure interface of liquid chromotography mass spectrometry, especially systems with atmospheric pressure ionization sources.

Factor 13 (DOUBLY, SINGLY, COULOMB, CHARGES, PROTONATION, INTRAMOLECULAR, ANIONS, PROTONS, ARGININE, ENERGY, CLUSTERS) – focuses on the influence of coulomb repulsion on the reaction and dissociation rates of singly and doubly charged ions (including protonated and deprotonated ions) within the mass spectrometry system.

Factor 14 (TRANSFER, REACTIONS, PROTONS, REACTIVITY, PROTONATION, NEUTRAL, GAS-PHASE, CHEMISTRY, ANIONS, BASES) – focuses on proton transfer reactivity in the gas-phase, and reaction of singly- and multiply-protonated molecules.

Factor 15 (SECTOR, MAGNETIC, INSTRUMENTS, RESOLUTION, HYBRID, EFFICIENCY) – focuses on mass spectrometric systems efficiency, based on magnetic sector instrumentation, for large protein mass and structure determination.

Factor 16 (ACETONITRILE, SOLVENTS, WATER, METHANOL, SOLUTIONS, AQUEOUS, ORGANIC) – focuses on the role of solvent compositions in the liquid and gel chromatographic separation process

Factor 17 (STORAGE, TRAP, QUADRUPOLE, TIME-OF-FLIGHT, INJECTION, SOURCE) – focuses on storage and accumulation of large source biopolymer ions in a quadrupole ion trap, and subsequent injection of these ions into the flight tube of a time-of-flight mass spectrometer. This process converts the typically continuous source ion beam into a higher density pulsed ion beam for the mass spectrometer, resulting in higher resolution and sensitivity.

Factor 18 (CARBOHYDRATE, HETEROGENEITY, OLIGOSACCHARIDES, GLYCOPROTEIN, STRUCTURES, MOIETY) – focuses on defining the structural heterogeneity of the carbohydrate oligosaccharide moiety of glycoprotein, using ionization mass spectrometry.

Factor 19 (AMMONIUM, ACETATE, BUFFER, PH, SALTS) – focuses on mass spectra of aqueous analyte solutions with varying concentrations of ammonium acetate,

Factor 20 (LABILE, EXCHANGE, LIGANDS, NMR, METHANOL, COMPLEXES, DERIVATIVES, ESI-MS) – focuses on determining exchange numbers and rates of labile protons (hydrogen/ deuterium exchange) and ligands by both electrospray ionization mass spectrometry and NMR.

Factor 21 (INTERNAL, LINEAR, STANDARDS, DYNAMICS, MONITORING, PLASMA, ASSAY, EXTRACTS) – focuses on the use of ionization mass spectrometry for the quantitative determination of organic compounds in plasma, making use of internal standards for quantification or standardization

Factor 22 (BOVINE, ALBUMIN, MYOGLOBIN, UBIQUITIN) – focuses on the fundamental generic sources of protein types and functions used in the determination of structural properties.

Thus, the 22 factors can be viewed as thrust areas constituting the lowest level taxonomy. There are myriad ways to combine the factors, depending on which dominant characteristics are chosen. A reasonable taxonomy is the following (cluster numbers are in parentheses, respective sub-cluster themes are in square brackets).

Separation (4, 16)
[chromotography, solvents]

Ionization Source/ Processes (2, 5, 8, 13, 14, 20)
[CID, droplet charge, alkali metal cations, coulomb repulsion, proton transfer, labile proton exchange]

Mass Analyzer (11, 12, 15, 17)
[ion cyclotron resonance, atmospheric pressure sources, magnetic sector instruments, quadrupole ion trap]

Mass Spectrometer System (3, 6, 21)

[MALDI, electrospray IMS, internal standards]

Applications Predominantly Biomedical (1, 7, 10, 22)
[amino acid structure, conformation, E Coli gene expression, generic proteins]

Applications Other, including some Biomedical (9, 18, 19)
[noncovalent complexes, carbohydrate structures, analyte solution mass spectra]

3.4.1.2. Multi-Link Clustering
A symmetrical co-occurrence matrix of the 253 highest frequency high technical content words was generated. The matrix elements were normalized using the Equivalence Index (Eij=Cij^2/Ci*Cj, where Ci is the total occurrence frequency of the ith phrase, and Cj is the total occurrence frequency of the jth phrase, for the matrix element ij), and a multi-link clustering analysis was performed using the WINSTAT statistical package. The Average Linkage method was used. A description of the final 253 phrase dendrogram (a hierarchical tree-like structure), and the aggregation of its branches into a taxonomy of categories, follows (See Figure 2 at end of report). The capitalized phrases in parentheses represent cluster boundary phrases for each category.The hierarchical structure of the taxonomy is guided by the branching structure of the dendrogram. As the dendrogram progresses from one hierarchical level to the next downward level, each branch divides into two parts. Thus, the highest level of the taxonomy consists of two clusters, the next level consists of four clusters, and so on. The hierarchical taxonomy described below follows the branching structure for the highest taxonomy levels, but aggregates the smaller clusters at the lower hierarchical level differently in some cases. For very small clusters, the algorithm will continue the sub-division process, sometimes resulting in physically unrealistic small clusters (e.g., one word). Essentially, the sub-division process along a given branch is terminated when the resulting clusters become artificially small.

[INSERT FIGURE 2]

At the lowest level of the present taxonomy hierarchy, the 253 phrases in the dendrogram are grouped into 24 clusters. Each cluster in this lowest hierarchical level is assigned a letter, ranging from A to X. Now, the hierarchical development of the taxonomy, and contents of each cluster, will be described.

In the dendrogram, the abcissa is the 253 words, and the ordinate is a measure of the cohesiveness of the words, or ‘distance’. The numerical range of the distance is from zero to about 520. If words, or word units, are linked at a low value of the distance,

their linking is strong. Many inherently double or triple word phrases will show component word links at very low distance values. As distance increases, the strength of the linkages progressively weakens. To aid in the interpretation of the dendrogram, the upper range of the distance was stretched, and the lower range was omitted from the dendrogram. This allows the critical branching at the highest hierarchical levels to be identified.

Starting from the phrase adjoining the 'distance' ordinate, the first main cluster (A-D) ranges from MASSES to COMPOUNDS. The second main cluster (E-X) ranges from CHARGES to BIOCHEMICAL. Cluster (A-D) is much smaller in extent and coverage than cluster (E-X).

The total dendrogram reflects different aspects of mass spectrometry techniques, phenomena, and applications. The first main cluster (A-D) covers different aspects of electrospray ionization mass spectrometry phenomena and applications at a high level of description. The second main cluster (E-X) addresses detailed phenomena (e.g., spray droplet charge density), applications (e.g., structure of bovine serum albumin), and complementary and competitive techniques to electrospray ionization (e.g., liquid chromatography, MALDI). As the factor matrix results showed, there are common phenomena and applications shared by the different mass spectrometry techniques. Therefore, these cluster structures should not be considered unique, but require consideration of the different perspectives provided by the multi-link and factor matrix approaches for completeness. Each of these large clusters will now be divided and sub-divided into smaller clusters, and discussed.

Because of the narrower coverage of cluster (A-D), it can be divided directly into its elemental clusters. Cluster A (MASSES to SOLUTIONS) focuses on the use of electrospray ionization mass spectrometry for the detection of mass spectra of large molecules in solution.

Cluster B (PROTEINS to WEIGHTS) focuses on characterization of protein structure and properties through molecular weight analysis of component peptide and amino acid building blocks using electrospray ionization mass spectrometry as a central technique.

Cluster C (FRAGMENTS to CLEAVAGES) focuses on the use of post-electrospray ionization collision-induced dissociation of macromolecules coupled with tandem mass spectrometry to analyze the ion mass spectra of the resulting fragments for structural analysis.

Cluster D (COMPLEXES to COMPOUNDS) focuses on the use of electrospray ionization mass spectrometry to study noncovalently bound ligands and complexes.

Cluster (E-X) can be divided into clusters (E-W) and X. Cluster (E-W) ranges from CHARGES to MOIETY, and cluster X ranges from AFFINITY to BIOCHEMICAL). Cluster (E-W) focuses on the applications and physical/ chemical phenomena of the electrospray and additional techniques in the Fenn citing papers. Cluster X focuses on the use of semi-automated techniques (incorporating electrospray ionization because of its relatively gentle nature) for high throughput screening of drugs that have high affinities to targets, typically high linked ligand affinities to protein macromolecular targets.

Cluster (E-W) can be divided into clusters (E-M) and (N-W). Cluster (E-M) ranges from CHARGES to OLIGOMERS, and cluster (N-W) ranges from LIQUID to MOIETY). Cluster (E-M) focuses mainly on the physics and chemistry of the sample separation process, ion generation and reaction process, and mass analysis phenomena and instrumentation. Cluster (N-W) focuses mainly on aggregate sample separation methods, MALDI characteristics, and biomedical applications primarily and organic chemistry applications secondarily.

Cluster (E-M) can be divided into clusters (E-K) and (L-M). Cluster (E-K) ranges from CHARGES to DETECTOR, and cluster (L-M) ranges from HYDROGEN to OLIGOMERS. The smaller cluster (L-M) separates out NMR and fast atom bombardment techniques. Because of the diversity of themes in clusters (E-K) and (L-M), both clusters will now be divided into their elemental clusters.

Cluster E (CHARGES to COULOMB) focuses on the generation of single and double ions during the MS process and, with the use of ion mobility techniques, conformation of peptide ions in the gas phase. Principal process for ion formation is through proton-transfer reactions.

Cluster F (METAL to CLUSTERS) focuses on use of negative ion mass spectrometry to study the influence of alkali metal salts on the clustering of ionic materials.

Cluster G (CONCENTRATIONS to METHANOL) focuses on the aqueous analyte solutions and media used in the separation and purification of amino acid and peptide residues by high pressure liquid chromatography and of the media generally used in the digestion of polypeptides and proteins.

Cluster H (SOURCE to CHARGE-STATE) focuses on phenomenological aspects of Fourier transform ion mass spectrometry, including cyclotron resonance excitation and quadrupole techniques, for high mass resolution MS instruments.

Cluster I (BOVINE to HEME) focuses on protein sources used in the current studies for mass spectrometric techniques and analysis.

Cluster J (RATES to CONE) focuses on the effect of electrostatic and liquid properties, and flow variables, on the size and charge of droplets ejected in conical sprays from capillaries for injection into the mass spectrometer.

Cluster K (INTERFACE to DETECTOR) focuses on the collision-activated dissociation at the vacuum/ atmospheric pressure interface between the high pressure liquid chromatography separation process and mass spectrometry ionization chamber.

Cluster L (HYDROGEN to RING) focuses on hydrogen exchange and labeling experiments coupled with methods to analyze protein backbone such as NMR, IR and X-ray crystallography to confirm protein structure. Additional structural information is derived through derivative formation and analysis of oxidation products.

Cluster M (ATOMS to OLIGOMERS) focuses on fast atom bombardment on proteins and oligomeric materials as an alternative ion source, and coupling the information to other spectroscopic techniques.

Again, because of the diversity of themes in cluster (N-W), this cluster will be divided into its elemental clusters.

Cluster N (LIQUID to APPLICATIONS) focuses on the use of reversed phase high pressure liquid chromatography and gel electrophoresis for the separation/purification of macromolecule residues and fragment samples prior to injection into an ionization mass spectrometer.

Cluster O (DIGESTS to ENZYME) focuses on peptide mapping and recombinantion of proteins to obtain structural and conformational information. Peptides are generated from a variety of digestion process including trypsin digests, separated typically in liquid or gel chromatography systems, and structurally analyzed by ionization mass spectrometry.

Cluster P (LASER to HYBRID) focuses on matrix-assisted laser desorption ionization time-of-flight mass spectrometry for peptide mass fingerprinting, followed by post source decay analysis for more detailed characterization of amino acid sequences.

Cluster Q (STANDARDS to CORRELATION) focuses on the use of ionization mass spectrometry for the quantitative determination of organic compounds in plasma, making use of internal standards for quantization, and identifying linear calibration curves and dynamic ranges.

Cluster R (BIOLOGICAL to TREATMENT) focuses on biological assays of DNA and DNA/RNA mimics, including sequencing techniques for oligonucleotides and DNA sizing analyses using synthetic oligonucleotides.

Cluster S (CELLS to INHIBITOR) focuses on analysis of phosphorylated peptides, especially phosphorylation at serrine residues in various cells, and mass determination of hydrophobic membranes.

Cluster T (NITROGEN to ARGININE) focuses on peptide analysis of nitrogen containing amino acids, lysine and arginine. Multiple ions formed during protonation of these residue bases and their subsequent analysis by mass spectrometric affords insights into the relation of side chain fragmentation to the residue's proximity to N-terminal or C-terminal groups.

Cluster U (PURIFICATION to POSTTRANSLATIONAL) focuses on purification to homogeneity of genetically expressed proteins in Escherichia Coli cells, followed by subsequent peptide sequencing showing removal of N-terminal methionine by posttranslational processing.

Cluster V (REACTIVITY to SUBSTRATE) focuses on protein structural and reactivity determination by mass spectrometry, especially disulfide bonding among cysteine residues.

Cluster W (COMPOSITION to MOIETY) focuses on defining the structural heterogeneity of the carbohydrate oligosaccharide moiety of glycoproteins, using ionization mass spectrometry.

3.4.1.3. Partitional Document Clustering

Document clustering is the grouping of similar documents into thematic categories. Different approaches exist [50-59]. The approach presented in this section is based on

a partitional clustering algorithm [60-61] contained within a software package named CLUTO. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO uses a randomized incremental optimization algorithm that is greedy in nature, and has low computational requirements. 32 individual clusters were chosen. Each Cluster is numbered (beginning with zero), and the number of documents in each cluster appears in parentheses at the beginning of every cluster. The most descriptive words (actually word stems) in each cluster are also shown in parentheses. Each word within the cluster is followed by a number that represents the percentage of intra-cluster similarity explained by the word. The theme of each cluster is represented by the initial high value keywords shown. The order of the clusters reflects the net cohesiveness (the intra-cluster similarity minus the inter-cluster similarity).

(29) Cluster 0 (jet 13.5, cone 8.8, drop 3.1, liquid 2.9, electr 2.6, flow 2.3, droplet 2.3, sprai 2.0, cone.jet 2.0, current 1.9, taylor 1.9, diamet 1.7, conduct 1.6, swirl 1.5, meniscu 1.5, breakup 1.4, surfac 1.2, size 1.1, taylor.cone 1.1, flow.rate 1.0)

(32) Cluster 1 (faim 8.2, conform 6.8, ion.mobil 5.8, mobil 5.3, cross.section 4.6, section 4.5, cross 3.7, ion 2.1, ga 1.3, charg 1.3, state 1.2, collis.cross.section 1.1, collis.cross 1.1, ga.phase 1.1, compact 0.9, charg.state 0.9, mobil.spectrometri 0.6, phase 0.6, separ 0.6, drift 0.6)

(27) Cluster 2 (atmospher 7.6, atmospher.pressur 7.1, pressur 5.0, apci 3.4, chemic.ioniz 2.5, pressur.chemic 2.3, atmospher.pressur.chemic 2.3, pressur.chemic.ioniz 2.0, interfac 1.8, sprai 1.4, atmospher.pressur.ioniz 1.4, pressur.ioniz 1.4, sampl 1.1, analyt 1.0, ion.sourc 1.0, ion 0.9, pesticid 0.9, liquid 0.8, api 0.8, chemic 0.8)

(34) Cluster 3 (protein 12.1, gel 7.1, databas 5.2, spot 3.1, sequenc 2.2, dimension 1.7, membran 1.7, maldi 1.6, peptid 1.5, two.dimension 1.4, electrophoresi 1.4, peptid.mass 1.2, search 1.0, gel.electrophoresi 1.0, protein.spot 0.9, stain 0.9, peptid.mass.fingerprint 0.8, digest 0.8, mass.fingerprint 0.7, separ 0.7)

(35) Cluster 4 (proton 11.1, charg 4.0, reaction 3.1, anion 3.0, proton.affin 2.5, proton.transfer 1.9, kcal 1.8, kcal.mol 1.8, multipli 1.8, ga.phase 1.8, transfer 1.7, barrier 1.6, ga 1.6, affin 1.4, state 1.3, mol 1.2, coulomb 1.2, repuls 1.2, phase 1.1, electron 1.1)

(43) Cluster 5 (droplet 26.8, charg 4.9, evapor 3.0, cluster 2.7, aerosol 2.2, ion.evapor 2.0, charg.droplet 1.7, particl 1.7, ion 1.3, surfac 1.3, size 1.2, mobil 1.1, diamet 0.9, solut 0.9, drop 0.8, dma 0.8, field 0.7, differenti.mobil 0.6, concentr 0.6, droplet.charg 0.5)

(33) Cluster 6 (capillari 11.9, electrophoresi 6.8, capillari.electrophoresi 5.4, cze 5.1, separ 3.7, ipg 2.3, zone 1.4, isoelectr 1.1, interfac 1.0, capillari.zone 0.9, coupl 0.9, protein 0.8, capillari.zone.electrophoresi 0.8, zone.electrophoresi 0.8, peptid 0.8, line 0.8, buffer 0.7, gradient 0.6, techniqu 0.6, focus 0.6)

(25) Cluster 7 (link 8.8, acyl 7.5, cross.link 6.0, asp 2.7, cross 2.7, toxin 2.1, subunit 1.9, asn 1.5, alpha 1.2, branch 1.2, c14 1.2, leu 1.1, beta 0.9, asp.asp 0.9, transducin 0.9, om 0.9, glucosyl 0.9, microcystin 0.8, glycopeptid 0.7, residu 0.7)

(36) Cluster 8 (proteom 10.8, technolog 5.8, protein 5.7, genom 5.5, function 2.7, advanc 1.5, vaccin 1.2, new 1.1, biolog 1.1, research 1.1, viral 1.0, tool 1.0, throughput 0.9, high.throughput 0.8, develop 0.7, diseas 0.7, recent 0.6, sequenc 0.6, biologi 0.6, structur.function 0.5)

(44) Cluster 9 (ligand 6.2, phosphin 6.1, cation 4.0, r2dtc 3.4, dtp 3.1, pph3 2.9, complex 2.6, esm 2.5, electrosprai.mass 1.8, solut 1.6, electrosprai.mass.spectra 1.0, nmr 1.0, mass.spectra 1.0, eta 0.9, eta2 0.9, angstrom 0.8, group 0.8, iii 0.7, spectra 0.7, bf4 0.7)

(60) Cluster 10 (charg 15.8, charg.state 10.9, state 8.6, ion 3.8, ion.ion 1.4, mass.charg 1.3, multipli 0.8, multipli.charg 0.8, proton 0.8, reaction 0.8, distribut 0.7, charg.ion 0.7, ion.ion.reaction 0.6, transfer 0.6, reduct 0.6, ion.reaction 0.6, proton.transfer 0.5, anion 0.5, ion.charg 0.5, spectra 0.5)

(50) Cluster 11 (maldi 9.4, matrix 2.4, tof 2.0, laser.desorpt 1.8, assist.laser.desorpt 1.8, assist.laser 1.8, matrix.assist.laser 1.8, laser 1.8, laser.desorpt.ioniz 1.8, matrix.assist 1.8, desorpt.ioniz 1.7, desorpt 1.6, assist 1.6, peptid 1.6, maldi.tof 1.6, time.flight 1.5, flight 1.5, oligosaccharid 1.4, flight.mass.spectrometri 1.2, desorpt.ioniz.time 1.2)

(38) Cluster 12 (pcr 9.2, dna 5.8, strand 5.1, pcr.product 3.7, lo 3.3, product 1.9, adduct 1.7, amplicon 1.2, doubl.strand 1.1, base 0.9, chiral 0.9, strand.dna 0.8, duplex 0.8, singl.strand 0.7, base.pair 0.7, pair 0.7, polymeras 0.7, str 0.7, cisplatin 0.6, oligonucleotid 0.6)

(46) Cluster 13 (exchang 10.9, conform 6.4, ga.phase 3.6, phase 3.1, hydrogen 3.0, ga 2.9, heme 2.6, state 1.8, charg 1.7, unfold 1.6, protein 1.4, denatur 1.2, deuterium 1.0, cytochrom 1.0, solut 0.9, ion 0.9, charg.state 0.8, disulfid 0.7, proton 0.7, stabil 0.7

(34) Cluster 14 (phosphoryl 15.5, cell 3.5, acp 3.1, express 2.6, cam 2.2, gene 2.1, kinas 2.0, protein 1.2, recombin 1.1, peptid 1.1, stimul 1.0, autophosphoryl 0.9, transcript 0.9, histon 0.8, site 0.8, regul 0.8, coli 0.8, apoc 0.8, serin 0.8, activ 0.7)

(43) Cluster 15 (beta 19.4, alpha 12.7, hemoglobin 3.3, globin 2.6, human 2.0, chain 1.4, gamma 1.0, esi 0.9, milk 0.9, lactamas 0.8, alpha.beta 0.8, subunit 0.8, glycat 0.8, beta.lactamas 0.7, variant 0.7, molecular 0.6, protein 0.5, lactosyl 0.5, enzym 0.5, adduct 0.5)

(46) Cluster 16 (ion 2.7, cyclotron 2.5, hexapol 2.3, fourier 2.1, fourier.transform 2.1, transform 2.0, fticr 1.7, ion.cyclotron 1.6, reson 1.6, cyclotron.reson 1.4, ion.cyclotron.reson 1.4, trap 1.2, transform.ion.cyclotron 1.2, transform.ion 1.2, fourier.transform.ion 1.2, icr 1.2, resolut 0.9, calibr 0.8, cyclotron.reson.mass 0.8, reson.mass 0.8)

(63) Cluster 17 (matrix 6.1, maldi 5.0, laser 4.7, desorpt 4.5, laser.desorpt 3.6, matrix.assist 3.0, assist.laser.desorpt 3.0, assist.laser 3.0, matrix.assist.laser 2.9, assist 2.7, desorpt.ioniz 2.4, laser.desorpt.ioniz 2.3, sampl 1.5, protein 1.4, method 1.0, polym 0.9, molecular 0.8, analyt 0.7, desorpt.ioniz.mass 0.7, prepar 0.6)

(39) Cluster 18 (column 5.2, digest 3.8, peptid 3.6, flow 2.7, tryptic 2.6, pack.capillari 1.6, pack 1.5, capillari 1.3, hplc 1.2, liquid.chromatographi 1.1, liquid 1.0, tryptic.digest 1.0, chromatographi 1.0, line 0.9, protein 0.9, separ 0.8, scan 0.8, system 0.7, sampl 0.7, compon 0.7)

(34) Cluster 19 (bind 17.1, dimer 4.2, ca2 3.7, calcium 3.5, librari 2.7, calmodulin 1.9, protein 1.5, dna 1.3, cooper 1.2, monom 0.9, esi 0.9, tetram 0.7, halophil 0.7, motif 0.6, bound 0.6, interact 0.6, mg2 0.6, compound 0.6, leucin.zipper 0.5, zipper 0.5)

(58) Cluster 20 (metal 20.9, complex 5.6, metal.ion 5.0, coordin 3.1, ligand 2.6, alkali 2.5, alkali.metal 2.2, ether 2.0, crown 1.6, solut 1.1, ion 1.1, metal.complex 0.9, supramolecular 0.9, transit.metal 0.8, bpy 0.8, crown.ether 0.8, cation 0.7, cluster 0.6, speci 0.6, alkali.metal.ion 0.6)

(38) Cluster 21 (mutant 4.8, site 2.8, proteas 2.3, protein 2.0, enzym 1.9, activ 1.8, mutagenesi 1.7, site.direct 1.5, bind 1.4, esim 1.4, mutat 1.3, residu 1.2, receptor 1.1, direct 1.1, cystein 1.0, domain 1.0, peptid 1.0, site.direct.mutagenesi 1.0, direct.mutagenesi 1.0, express 0.9)

(37) Cluster 22 (capillari 5.7, flow 2.3, min 2.1, tip 2.0, funnel 1.9, ion.funnel 1.7, needl 1.5, sprai 1.4, solut 1.3, transmiss 1.2, flow.rate 1.1, ion 1.0, voltag 1.0, current 0.9, gold 0.9, fluoresc 0.9, interfac 0.8, micel 0.8, rate 0.8, esi 0.7)

(45) Cluster 23 (complex 8.9, esi 4.6, noncoval 3.0, interact 3.0, protein 2.9, coval 2.5, non.coval 2.2, rna 2.1, bind 1.7, inhibitor 1.7, stoichiometri 1.6, non 1.1, peptid 0.9, enzym 0.9, protein.complex 0.9, noncoval.complex 0.8, affin 0.8, drug 0.8, tar 0.7, electrosprai.ioniz.mass 0.7)

(61) Cluster 24 (fragment 5.6, cid 4.4, collis 4.2, energi 4.0, sid 3.0, induc.dissoci 2.3, dissoci 2.2, proton 1.7, induc 1.7, peptid 1.6, ion 1.6, collis.induc 1.1, collis.induc.dissoci 1.1, low.energi 0.9, fragment.ion 0.8, methyl 0.7, cleavag 0.7, tandem.mass 0.6, tandem 0.6, structur 0.5)

(57) Cluster 25 (dissoci 8.9, ion 3.2, energi 1.9, proton 1.8, product.ion 1.6, loss 1.6, charg 1.6, fragment 1.5, product 1.4, cluster 1.3, activ 1.1, nucleobas 1.1, base 1.0, collision.activ 0.9, collision 0.8, sequenc 0.8, oligonucleotid 0.7, bond 0.7, mer 0.7, oligom 0.6)

(45) Cluster 26 (ion.mode 3.6, neg 3.3, mode 2.9, neg.ion 2.7, acid 2.1, esi 2.0, posit 2.0, ion 2.0, compound 1.5, mobil.phase 1.5, posit.ion 1.4, posit.ion.mode 1.2, anion 1.1, neg.ion.mode 1.1, fatti.acid 0.9, fatti 0.8, solut 0.7, salt 0.7, mobil 0.7, unsatur 0.6)

(44) Cluster 27 (hplc 3.1, oligosaccharid 2.7, method 1.7, tandem 1.6, esi 1.6, quantit 1.6, plasma 1.4, standard 1.4, chromatographi 1.4, liquid.chromatographi 1.3, sampl 1.3, label 1.3, extract 1.3, tandem.mass 1.3, liquid 1.1, detect 1.0, odn 1.0, phospholipid 1.0, intern.standard 0.9, tandem.mass.spectrometri 0.9)

(61) Cluster 28 (trap 6.7, ion 4.8, ion.trap 4.1, instrument 2.0, spectromet 1.8, mass.spectromet 1.7, quadrupol 1.7, resolut 1.2, detector 1.1, time 0.9, sector 0.9, sourc 0.8, flight 0.8, time.flight 0.7, puls 0.6, storag 0.6, high 0.6, magnet 0.6, energi 0.6, tof 0.6)

(51) Cluster 29 (reaction 4.7, intermedi 3.8, esi 1.6, solut 1.5, radic 1.5, electron 1.4,

oxid 1.4, speci 1.4, solvent 1.3, lithium 1.3, fulleren 1.2, compound 1.1, molecul 1.1, kinet 1.0, radic.cation 1.0, complex 0.9, spectroscopi 0.7, detect 0.7, reagent 0.7, lithium.ion 0.7)

(92) Cluster 30 (sequenc 5.5, peptid 5.2, amino 4.1, amino.acid 3.7, protein 3.2, residu 2.6, acid 2.3, termin 1.9, amino.acid.sequenc 1.9, acid.sequenc 1.8, disulfid 1.3, molecular 1.0, purifi 0.8, modif 0.8, weight 0.7, molecular.weight 0.7, cystein 0.6, ident 0.6, molecular.mass 0.6, methionin 0.6)

(51) Cluster 31 (applic 2.7, method 2.1, analyt 2.0, techniqu 1.6, chromatograph 1.3, drug 1.3, mass.measur 1.2, chromatographi 1.2, materi 0.9, liquid 0.8, pharmaceut 0.8, compound 0.7, sugar 0.7, liquid.chromatographi 0.7, line 0.6, degrad.product 0.6, tool 0.6, calibr 0.6, sampl 0.6, character 0.6)

The CLUTO algorithm then aggregates the clusters in a hierarchical taxonomy. The high value words extracted in each category are shown in parentheses. Overall, the main category (ionization, protein, peptide, charge, ESI, complex, sequence, acid), Level 1, contains 1431 records, with a broad focus of bio-molecular applications and the ionization-charge components of the mass detection and analysis process. Level 2 contains the first major categorical split of two categories: Applications and Ionization Process. There are 532 records in Applications (protein, peptide, sequence, MALDI, binding, DNA, acid, amino), focused on large bio-molecules. Additionally, there are 899 records in Ionization Process (ionization, charge, proton, solutions, electrospray ionization, state, fragment, dissociation), focusing on the charging process and charge state, as well as the sample solution prior to ionization.

Level 3 contains the next categorical split of 4 categories: Bio-molecule Structure, MALDI Protein Mapping, Ionization, and Sample Preparation. The Applications category sub-divides into Bio-molecule Structure and MALDI Protein Mapping. There are 349 records in Bio-molecule Structure (protein, peptide, binding, sequence, residue, beta, alpha, amino), focused on proteins, peptides, binding states, and amino acid sequencing.

There are 183 records in MALDI Protein Mapping (MALDI, protein, matrix, laser, desorption), focused on the use of MALDI for protein mapping. Sampling of these records shows the main focus to be MALDI, with Fenn/ ESI appearing mainly as a reference. Appearance of MALDI papers in the Fenn citing papers implies that either ESI is being cited as a MALDI alternative for Protein Mapping or that ESI is being cited historically as a demonstration that large bio-molecule mass measurements were

possible.

Who are the MALDI researchers most cited in the Fenn citing papers? As Table 5B (most cited authors) shows, the main soft laser desorption researchers listed are Karas/ Hillenkamp. Tanaka does not appear in the top twenty list. To test whether this result applies beyond the Fenn citing papers, in a more recent context, a database of 300 papers was generated from the SCI. The query used was the same as in the Background (laser and desorption and (ion* or mass spectrometry)), and the records were the most recent prior to October 2002 (so as not to be influenced by the Nobel awards). After the elimination of (few) self-citations, the citation results were as follows: Karas-70 citations; Hillenkamp-25 citations; Tanaka-18 citations; Beavis-12 citations. 79% of the Karas citations were pre-1989 (1985-1988). These results mirror those using MALDI as the query term. Remembering that the SCI provides the first author in citation print-outs, and most of the early soft laser desorption papers of Karas and Hillenkamp were joint, it appears that the early works most referenced on soft laser desorption/ MALDI are those of Karas/ Hillenkamp. As shown in the Background, it was true over a decade ago, and as shown in this paragraph, it remains true today.

The Ionization Process category sub-divides into Ionization and Sample Preparation. There are 398 records in Ionization (ionization, charge, proton, charge state, dissociation, energy, fragment), focused on characteristics of the charged state. There are 501 records in Sample Preparation (droplet, solution, metal, ion, capillary, complex, liquid), focused on the process and components preparatory to ionization.

At a taxonomy level of about six (~30-40 clusters), the number of document clusters is about the same order as the number of factors or number of word/ phrase clusters. The themes of the document clusters at this level are similar to the themes of the factors or word/ phrase clusters. Because of the aggregation methodology used in CLUTO, the higher level taxonomy categories are easier to define than with the factors or word/ phrase clusters. Additionally, the document clustering output shows the number of records in each document cluster and each taxonomy category at every level, thereby allowing estimates of level of effort to be made for each category.

3.4.2. Tanaka Citing Papers

The words contained in the Tanaka citing paper Abstracts were extracted by the Vantage Point software, and sorted by frequency of occurrence. The highest frequency high technical content words were identified by inspection. Very similar

words were consolidated (e.g., singulars/ plurals, full spellings/ acronyms, very strong synonyms).

3.4.2.1. Factor Matrix Clustering

A correlation matrix of the 253 resultant consolidated words was generated, and a factor analysis was performed using the WINSTAT statistical package (an Excel add-in). The eigenvalue floor was set equal to unity to insure that each resulting factor provide value-added information, and a factor matrix consisting of 51 factors resulted. Due to space limitations, only the larger factors (1-24) will be analyzed. A description of each factor, and the aggregation of all factors into a taxonomy, follows.

Factor 1 (SEQUENCING, DIGESTION, DISULFIDE, PEPTIDES, MAPPING, TRYPTIC, AMINO, RESIDUES, PRIMARY, PROTEINS, STRUCTURES, NATIVE, MULTIPLE, ENZYME, REDUCTION, BONDS, ISOLATED, HPLC, SITES, ACIDS) – focuses on characterization of proteins' structures and properties through their component peptide structures by molecular weight determination and peptide amino acid sequences and residues, using MALDI mass spectrometry as a central technique. It includes peptide mapping of recombinant proteins to obtain structural and conformational information. Peptides are generated by protein digestion and typically separated by high pressure liquid chromatography.

Factor 2 (MASSES, LASER, SPECTROMETRY, MATRIX-ASSISTED, TIME-OF-FLIGHT, IONS, MALDI, DESORPTION/ IONIZATION, SPECTRA, MATRIX, DESORPTION, METHOD, MOLECULAR, FLIGHT, RESOLUTION, WEIGHTS, DETECTION, POLYMERS, ANALYTE) – focuses on using matrix-assisted laser desorption ionization mass spectrometry with time-of-flight mass analyzer for molecular weight determination, with emphasis on polymers.

Factor 3 (NMR, CYCLIC, SYNTHESIZED, GROUPS, GPC, CHAINS, OLIGOMERS, COMPOSITION, POLYMERIZATION, OXIDE, REACTIONS, POLYMERS, WEIGHTS, SERIES, ETHYLENE, STRUCTURES, SPECTROSCOPY, MS, MOLECULAR, CHROMOTOGRAPHY, LOSS, DITHRANOL, TERMINAL, PROPERTIES) – focuses on characterizing oligomer and polymer average molecular weights from MALDI, and comparing against NMR and GPC analysis. Emphasis is on cyclic oligomers, and end groups of synthetic polymers, especially on the relation between the composition of the terminal group on a polymer chain and the ion yields. Synthetic polymers such as polystyrene and poly (ethylene glycol) are emphasized

Factor 4 (CYCLOTRON, FOURIER, TRANSFORM, RESONANT, TRAPPING, EXCITATION, SOURCES, POWER, COOLING, FTMS, SIGNAL-TO-NOISE, RESOLVING, ANALYZING, POTENTIALS, EFFICIENCY, M/Z, QUADRUPOLE, MAGNETIC, GRAMICIDIN, VELOCITY, RESOLUTION) – focuses on laser desorption Fourier transform ion cyclotron resonance mass spectrometry, concentrating on cooling and axializing the MALDI-generated ions by azimuthal quadrupolar excitation in the presence of collisions with neutral atoms in the source compartment of a dual ion trap, followed by detection of the axialized ions at much low pressure and much higher mass resolving power.

Factor 5 (HYDROXYCINNAMIC, ALPHA-CYANO-4, POSITIVE-ION, NEGATIVE-ION, 2.5 DIHYDROXYBENZOIC, MAGNETIC, HYDROGEN, DERIVATIVES, CARBON, ACIDS, LOSS, STRONG, SINAPINIC, SALTS, ADDUCT, MATRIX, WATER, ALKALI, OLIGOSACCHARIDES, REDUCTION) – focuses on identifying the spectra complexity and signal strength differences in positive-ion and negative-ion modes. Emphasis is on use of the matrices alpha-cyano-4-hydroxycinnamic acid and 2,5-dihydroxybenzoic acid, including analysis of the strong dependence of fragmentation on the nature of the matrix and on the presence or absence of water in the matrix solvent.

Factor 6 (YAG, THIN, ND, GRAMICIDIN, LDI, SUBSTRATE, LIGHT, DEPOSITION, VACUUM, WAVELENGTHS, SOLVENT, PARTICLES, ABLATION, PLATE FLUENCE, CLUSTERS, SILVER) – focuses on deposition of organic samples on stable thin precious metal film substrates, to minimize background interference with analyte ion peaks, facilitate deposition of samples from a variety of solvent systems, and generate analyte adduct ions in some cases. Emphasis is on desorption and ionization by Nd:YAG laser irradiation, and maximizing light absorption by film thickness variation.

Factor 7 (BOMBARDMENT, ATOMS, FAB, SIMS, ESI, LIQUID, PRIMARY, MS/MS, SECONDARY, STRUCTURES) – focuses on use of fast atom bombardment mass spectrometry analysis.

Factor 8 (DISSOCIATION, COLLISION, TANDEM, FRAGMENTS, ENERGY, ACTIVE, GAS, QUADRUPOLE, FAB, MODEL) – focuses on the use of tandem mass spectrometry with collision-induced dissociation to provide structural information for unknown sample molecules.

Factor 9 (UV-MALDI, INFRARED, WAVELENGTHS, ULTRAVIOLET, WATER, OPTICAL, YAG, TRANSITION, TRANSFER, GLYCEROL, IRRADIATION) – focuses on direct comparisons of the effectiveness of UV and IR lasers in MALDI mass

spectrometry, especially YAG lasers, and concentrates on the role of water in the matrix-analyte as a major laser energy absorber.

Factor 10 (FOCUSING, REFLECTING, PULSES, FIELDS, REFLECTRON, BROAD, METASTABLE, RESOLUTION, TIME-OF-FLIGHT, PLATE, VOLTAGE, ACCELERATION) – focuses on increasing the resolving power of time-of-flight spectra, using delayed extraction of MALDI-generated ions and high accelerating voltage reflecting mirror fields to focus the ions and extend their flight paths.

Factor 11 (CAPILLARY, GEL, SEPARATION, COUPLING, CHROMATOGRAPHY, TRYPTIC, CONCENTRATIONS, DIGESTION, CELLS) – focuses on combination of capillary and gel permeation chromatography to separate materials for subsequent injection into the mass spectrometer.

Factor 12 (MULTIPLY, SINGLY, POTENTIALS, ESI, SPECIES, WEIGHTS) – focuses on the formation of singly and multiply charged molecular ions via the field-assisted ion evaporation mechanism during electrospray ionization.

Factor 13 (MIRROR, VOLTAGE, FIELDS, INSULIN, BOVINE, ACCELERATION, POWER, SHOT, OPTICAL, RESOLVING, EXTRACTION) – focuses on extending the flight path by the use of reflecting electrostatic mirrors to increase the time-of-flight of MALDI molecular ions for increased mass resolution, concentrating on delayed extraction from high acceleration voltage sources.

Factor 14 (SODIUM, ADDUCT, PROTONATED, CATIONS, SPECIES, PEAKS, LIMITS, COMPOUNDS, ACIDS, STANDARDS) – focuses on use of counter ions that accompany much of the digestion process used in peptides and macromolecules prior to introduction into an ion mass spectrometer

Factor 15 (GAS-PHASE, AFFINITY, CHEMICAL, STATES, HYDROGEN, ELECTRON, METAL, REDUCTION, AGENT, OXIDE, REACTIONS, TRANSFER, CATIONS, LDI, THERMALLY) – focuses on the effects of proton affinity of MALDI matrices on the relative protonation of analytes from radical matrix molecular ions or protonated matrix ions.

Factor 16 (AGENT, GRAPHITE, RADIATION, MATERIALS, SYNTHETIC, PLATE, TRANSFER, OLIGOSACCHARIDES, STEP, POLAR, SPECIES, SPOT) – focuses on the role of carbon and other metallic materials as a matrix target plate component,

acting as an energy transfer agent by enhanced radiation absorption, and enhancing desorption of solvent and analyte ions.

Factor 17 (FWHM, MOLAR, PEAKS, SINAPINIC, OPTICAL, PH, PROTONATED, METASTABLE, BOVINE, ADDUCT) – focuses on a combination of substrate materials (sinapinic acid, bovine) with the optical processes for MALDI analysis of large macromolecules, concentrating on the FWHM intensity and breadth of protonated peaks.

Factor 18 (SOLUTIONS, SHOT, WATER, ANALYTE, GLYCEROL, COMPOUNDS, STABLE) – focuses on the use of analyte and matrix aqueous solutions for MALDI, especially water and glycerol matrix additions, and the resultant ion signal stability for repeated laser shots.

Factor 19 (TRANSITION, PHASE, VELOCITY, GAS, PROPERTIES, CO, STATES) – focuses on the effect of solid-to-gas phase transition from laser irradiation on subsequent ionization and lift-off velocity of analyte molecules.

Factor 20 (EJECTED, FLUENCE, ABLATION, VACUUM, MODEL, MOLECULES, EFFICIENCY) – focuses on modeling the ejection of analyte molecules at different laser fluences, concentrating on the dependence of the yield on fluence near ablation threshold to distinguish between ejection models.

Factor 21 (NEGATIVE, KINETIC, SIGNAL-TO-NOISE) focuses on solution degradation kinetics, emphasizing negative-ion mode operation with deprotonated molecular ions for high signal-to-noise mass spectra.

Factor 22 (POLY(ETHYLENE), ETHYLENE, OLIGOMERS) focuses on characterizing average molecular weights of oligomers, and end groups of synthetic polymers using MALDI. Synthetic polymers such as polystyrene and poly (ethylene glycol) are emphasized.

Factor 23 (DITHRANOL, SILVER, SALTS, CATIONS, POLYMERS, SYNTHETIC) focuses on the MALDI analysis of synthetic polymers, especially polystyrene, with use of dithranol as a matrix and silver salt ions as dopant to enhance the cationization of polystyrene through the formation of adduct complexes.

Factor 24 (BONDS, COMPLEXES, HYDROPHOBIC, SITES) focuses on determining the interaction strength in non-covalently bound complexes, and the influence of

hydrophobic interactions to establish differences between solution-phase and gas-phase binding energies.

Thus, the 24 factors can be viewed as thrust areas constituting the lowest level taxonomy. There are myriad ways to combine the factors, depending on which dominant characteristics are chosen. A reasonable taxonomy is the following (cluster numbers are in parentheses).

Separation (11)
[capillary/ gel chromatography]

Sample Preparation (5, 6, 16, 18, 23)
[alpha-cyano-4-hydroxycinnamic matrices, thin film substrates, carbon/ metallic matrices, aqueous matrix solutions, dithranol matrices/ silver salts dopants]

Ionization source/ process (7, 8, 9, 12, 14, 15, 19, 20)
[fast atom bombardment, CID, UV-infrared, multiply-charged ESI, protonated alkali adducts, gas-phase affinity, phase transition, analyte molecule ejection]

Mass Analyzer (4, 10, 13, 17, 21)
[ion cyclotron resonance, accelerating voltage focusing, reflecting mirror fields, signal processing FWHM, high S/N negative ion mode]

Mass Spectrometer System (2)
[MALDI]

Applications Predominately Biomedical (1)
[amino acid sequencing]

Applications Other (3, 22, 24)
[cyclic oligomer and synthetic polymer characterization, synthetic polymer characterization, covalent complex interactions]

3.4.2.2. Multi-Link Clustering

A symmetrical co-occurrence matrix of the 253 highest frequency high technical content words was generated. A description of the final 253 phrase dendrogram, and the aggregation of its branches into a taxonomy of categories, follows (See Figure 3 at end of report). At the lowest level of the present taxonomy hierarchy, the 253 phrases in the dendrogram are grouped into 29 clusters. Each cluster in this lowest hierarchical

level is assigned a letter, ranging from A to AC. Now, the hierarchical development of the taxonomy, and contents of each cluster, will be described.

Starting from the phrase adjoining the ‘distance’ ordinate, the first main cluster (A-B) ranges from MASSES to COMBINED. The second main cluster (C-AC) ranges from PROTEINS to METASTABLE. Cluster (A-B) is much smaller in extent and coverage than cluster (C-AC).

Cluster (A-B) is a high-level generic description of the common theme of Tanaka’s’s citing papers, namely, the MALDI approach, and would be expected to have a strong focus. The second cluster (C-AC) is a more detailed description of MALDI technique components, associated physical and chemical phenomena, and different MALDI applications. However, the citing literature does not separate applications from detection techniques and physiochemical phenomena as cleanly as reported in the more fundamental literatures. As the factor matrix results showed, there is considerable overlap and cross-linkage among these different types of clusters, negating the possibility of a purely orthogonal taxonomy. Therefore, these cluster structures should not be considered unique, but require consideration of the different perspectives provided by the multi-link and factor matrix approaches for completeness. Each of these large clusters will now be divided and sub-divided into smaller clusters, and discussed. At the higher and broader taxonomy levels, the divisions follow the branching of the dendrogram. At the lower and more detailed taxonomy levels, some branches are aggregated into clusters to eliminate excess fragmentation.

Because of the limited size of cluster (A-B), it can be divided directly into its elemental clusters. Cluster A (MASSES to ACIDS) focuses on using matrix-assisted laser desorption ionization mass spectrometry with time-of-flight mass analyzer for molecular weight determination, emphasizing hydrocarbon polymers with organic acid matrices.

Cluster B (METHOD-COMBINED) focuses on high resolution detection of analyte molecules and fragments, varying mixtures of compounds and laser wavelength-matrix combinations to maximize output signal intensity.

Cluster (C-AC) can be divided into clusters (C-J) and (K-AC). Cluster (C-J) ranges from PROTEINS to OLIGONUCLEOTIDES, and cluster (K-AC) ranges from COMPOUNDS to METASTABLE). Cluster (C-J) focuses on structural determination of proteins using MALDI and other ionization mass spectrometry techniques. Cluster (K-AC) focuses on the effect of different matrix materials on the analyte ionization spectra,

as well as chemical and physical pathways, and associated kinetic-internal energy transfers, for ion formation, and provision of detailed structural information.

Cluster (C-J) can be divided into clusters (C-I) and J, and cluster (K-AC) can be divided into clusters (K-T) and (U-AC). Cluster (C-I) ranges from PROTEINS to CYTOCHROME, and cluster J ranges from ATOMS to OLIGONUCLEOTIDES. Cluster (C-I) focuses mainly on structural determination of proteins using MALDI, with some corollary emphasis on structural determination of synthetic polymers. Cluster J focuses on comparison of fast atom bombardment and secondary ion mass spectrometry for protein analysis. Cluster (K-T) ranges from COMPOUNDS to EJECTED, and cluster (U-AC) ranges from ENERGY to METASTABLE. Cluster (K-T) focuses on the effect of different matrix materials on the analyte ionization spectra, and subsequent resolution accuracy of final mass spectra. Cluster (U-AC) focuses on chemical and physical pathways, and associated kinetic-internal energy transfers, for ion formation, and provision of detailed structural information.

All the elemental clusters from C to AC will now be described. Cluster C (PROTEINS-NATIVE) focuses on characterization of proteins' structures and properties through their component peptide structures by molecular weight determination and peptide amino acid sequencing, using MALDI mass spectrometry as a central technique. It includes peptide mapping of recombinant proteins to obtain structural and conformational information. Peptides are generated from trypsin digestion, separated typically by high pressure liquid chromatography, and structurally analyzed.

Cluster D (BONDS-AFFINITY) focuses on dynamic regulation of protein complexes in membrane proteins, emphasizing the impact of binding site and substrate affinity.

Cluster E (OXIDE-PURIFIED) focuses on structural characterization of terminal groups, in peptides, and the effect of enzymatic cleavage of C- or N-terminal peptides on subsequent biological activity and ionic signal intensity.

Cluster F (ANALYZING-BIOLOGICAL) focuses on use of MALDI to probe the biological activity within cells.

Cluster G (SINGLE-COMPONENTS) focuses on altering production of singly or multiply charged fragments resulting from MALD ionization, especially the lower molecular weight fragments such as CO.

Cluster H (POLYMERS-SPECTROSCOPY) focuses on characterizing average molecular weights, oligomer repeat units, and terminal group analysis of synthetic polymers using MALDI coupled with NMR spectroscopy. Synthetic polymers such as polystyrene and poly (ethylene glycol) are emphasized.

Cluster I (MS-CYTOCHROME) focuses on a combination of capillary GPC and gel electrphoresis separation of protein or synthetic polymer residues with MALDI for (mainly) structural characterization and identification.

Cluster J (ATOMS-OLIGONUCLEOTIDES) focuses on use of fast atom bombardment for protein analysis, and comparison of results with those obtained by secondary ion mass spectrometry.

Cluster K (COMPOUNDS-INORGANIC) focuses on structural analysis of thermally labile, polar and high molecular weight organic and inorganic compounds, emphasizing the use of nitrogen lasers for MALDI.

Cluster L (SURFACES-SHOT) focuses on the target matrix environment, emphasizing the performance and suitability for time-of-flight mass spectrometry of different particle materials and sizes, suspended in a variety of different liquids, especially glycerol-based.

Cluster M (MATERIALS-AGENT) focuses on the role of carbon and other metallic materials as a matrix target plate component, acting as an energy transfer agent by enhanced radiation absorption, and enhancing desorption of solvent and analyte ions.

Cluster N (CATIONS-SILVER) focuses on metal alkali ionization of neutral species with sodium or potassium ions, allowing non-biomolecular analytes to be detected as cation adducts and, in the case of proteins, as protonated molecular ions. Additionally, there was a focus on the formation of an adduct complex of metal cations and polystyrene, where the metal salt was observed in the mass spectra as well. This cluster emphasizes the use of dithranol as the matrix, and silver or copper ions as the dopanrt cations, for intense interactions between cation and polymer for MALDI.

Cluster O (PEAKS-LOSS) focuses on production of protonated molecular species from proteins in a sinapinic acid matric, and relating adduct-formation-driven peak broadening to parameters such as absorber concentration and laser irradiance.

Cluster P (CONCENTRATIONS-SIGNAL-TO-NOISE) focuses on accurate determination of component concentrations, using internal standards for fast, sensitive, and reproducible quantification. The impact of positive and negative ion operational modes on signal-to-noise ratio and concentration determination accuracy was emphasized.

Cluster Q (HYDROGEN-DERIVATIVES) focuses on analyte protonation and deprotonation using MALDI and magnetic sector analyzers, and concentrates on identifying the spectra complexity and signal strength differences in positive-ion and negative-ion modes. Emphasis is on use of the matrix materials alpha-cyano-4-hydroxycinnamic acid and 2,5-dihydroxybenzoic acid, including analysis of the strong dependence of fragmentation on the nature of the matrix and on the presence or absence of water in the matrix solvent.

Cluster R (ULTRAVIOLET-DIRECT) focuses on direct comparisons of the effectiveness of UV and IR lasers in MALDI mass spectrometry.

Cluster S (DEPOSITION-SOLVENT) focuses on deposition of organic samples on stable thin metal film substrates, to minimize background interference with analyte ion peaks, facilitate deposition of samples from a variety of solvent systems, and produce resolution near the detection limit. Emphasis is on desorption and ionization by Nd:YAG laser irradiation, and maximizing light absorption by film thickness variation.

Cluster T (MODEL-EJECTED) focuses on modeling the ejection of analyte molecules at different laser fluences, concentrating on the dependence of the yield on fluence near ablation threshold to distinguish between ejection models, and on the relationship between the depth of origin of the analyte ions and the threshold dependence of the yield.

Cluster U (ENERGY-GAS-PHASE) focuses on the chemical and physical pathways involved in MALDI ion formation, concentrating on analyte laser vaporization followed by multi-photon ionization in the gas phase to molecular radical ions. Emphasis is on internal and kinetic energy transfers, and their relation to reaction and ionization rates, including proton transfer from the ground state protonated matrix ions to MALDI analytes.

Cluster V (PHASE-PROPERTIES) focuses on MALDI laser heating of the matrix up to the phase transition temperature, and subsequent gas phase analysis of velocity distributions of molecules.

Cluster W (ESI-MS/MS) focuses on the use of tandem mass spectrometry, where the first quadrupole MS filters the analyte solution for materials of interest only, then subsequent collisions are used to fragment these analytes, and their daughter ions swept into a second time-of-flight MS where they are separated and analyzed.

Cluster X (SOURCES-POWER) focuses on laser desorption Fourier transform ion cyclotron resonance mass spectrometry, concentrating on cooling and axializing the MALDI-generated ions by azimuthal quadrupolar excitation in the presence of collisions with neutral atoms in the source compartment of a dual ion trap, followed by detection of the axialized ions at much low pressure and much higher mass resolving power.

Cluster Y (WORK-CURRENT) focuses on liquid matrices that prolong the analyte ion current without moving the laser irradiation spot and provide an ion current that is stable enough to acquire both mass spectra and collision-induced dissociation (CID) spectra.

Cluster Z (TARGET-SCANNING) focuses on electron scanning microscopy to determine homogeneity of target samples, as a function of Ph.

Cluster AA (LINEAR-FWHM) focuses on the use of MALDI linear time-of-flight mass spectrometry for analysis of large biomolecules, especially bovine insulin, with uniform distribution of matrix/analyte microcrystals over the entire sample surface as observed by reflection optical microscopy. Some emphasis was placed on the influence of the matrix/analyte molar ratio on the molecular ion yield.

Cluster AB (PULSES-ELECTROSTATIC) focuses on mass spectra resulting from sample spot exposure to laser irradiation, varying number of laser pulses, laser fluence, sample thickness, matrix-to-analyte molar ratio, total deposited amount, and analyte molecular mass. Additionally, the cluster focuses on increasing the resolving power of time-of-flight spectra, using delayed extraction of MALDI-generated ions and high accelerating voltage reflecting mirror fields to focus the ions and extend their flight paths.

Cluster AC (DECAY-METASTABLE) focuses on product ion mass spectra of ions formed by meta-stable or post-source decay, allowing structural information on analyte molecules to be obtained by extraction of sequence or substituent information on e.g., individual peptides contained in an enzymatic digest.

3.4.2.3. Partitional Clustering

The cluster display structure is the same as in the Fenn analysis section.

(9) Cluster 0 (ethoxyl 4.9, sim 3.8, tof.sim 3.4, polym 2.1, molecular.weight 1.3, weight 1.3, low.molecular.weight 1.2, tof 1.2, low.molecular 1.1, tof.maldi 1.0, surfynol 0.9, low 0.8, oligom 0.7, ethoxyl.polym 0.6, segreg 0.5, molecular 0.5, maldi 0.4, molecular.weight.compound 0.4, weight.compound 0.4, cation 0.4)

(9) Cluster 1 (subunit 7.8, alpha 2.0, gamma 1.9, pde 1.5, trbk 1.5, cross.link 1.5, membran 1.4, pgamma 1.2, gene 1.1, link 1.1, cross 0.9, residu 0.8, trbc 0.8, terminu 0.8, alpha.subunit 0.7, mutant 0.6, plasmid 0.6, rhodopsin 0.6, cy 0.6, pilin 0.5)

(9) Cluster 2 (sequenc 4.2, amino 2.5, amino.acid 2.3, peptid 2.1, amino.acid.sequenc 1.6, acid.sequenc 1.6, azurin 1.6, protein 1.4, profilaggrin 0.8, filaggrin 0.8, acid 0.5, mavicyanin 0.5, site 0.5, nucleas 0.5, protein.peptid 0.5, descriptor 0.4, protamin 0.4, modif 0.4, primari 0.4, column 0.4)

(9) Cluster 3 (tryptic 2.4, cy 1.6, maxadilan 1.6, linkag 1.5, digest 1.3, fab 1.3, peptid 1.2, tryptic.digest 1.1, max 1.0, imac 0.9, esi 0.8, contain.peptid 0.8, rhgh 0.7, histidin 0.6, cy.cy 0.6, charg 0.6, column 0.5, strategi 0.5, polyamin 0.5, im2 0.4)

(10) Cluster 4 (proton 6.1, transfer 3.2, proton.transfer 2.7, radic 2.4, radic.cation 1.4, analyt 1.1, proton.affin 1.0, matric 1.0, molecular.radic 0.8, cation 0.8, chemic.ioniz 0.6, molecular.radic.cation 0.6, non.polar 0.5, transfer.ioniz 0.5, affin 0.5, matrix.activ 0.5, ortho 0.5, charg.transfer 0.4, ground.state 0.4, cluster 0.4)

(10) Cluster 5 (detector 2.0, ion 1.3, acceler 1.2, resolut 1.1, dynod 0.8, plate 0.8, signal 0.8, microchannel 0.8, microchannel.plate 0.8, ion.detector 0.7, puls 0.6, ion.speci 0.6, speci 0.6, ion.signal 0.5, puls.focus 0.5, initi 0.5, mass.resolut 0.4, high.resolut 0.4, direct.matrix 0.4, new.instrument 0.4)

(9) Cluster 6 (bind 2.8, complex 2.6, p55 2.4, cam 2.3, interact 1.9, ica 1.8, templat 1.5, ca2 1.3, gpc.p55 1.0, sulfon 0.9, guest 0.9, gpc 0.9, metal 0.9, ligand 0.8, dimer 0.7, phase 0.7, noncoval 0.6, ica.ion 0.6, rhenium 0.6, rhenium.complex 0.6)

(10) Cluster 7 (antibodi 5.3, monoclon.antibodi 2.1, antigen 2.1, monoclon 1.9, thiol 1.9, disulfid 1.6, gnrh 1.5, disulfid.bond 1.4, epitop 1.3, secret 1.2, polypeptid 1.1, react 0.9, tunic 0.8, rhoph3 0.8, bond 0.7, cystein 0.6, golgi 0.6, mab 0.6, human 0.5, chain

0.5)

(9) Cluster 8 (label 3.5, assembl 1.0, probe 1.0, c36h6 0.8, sampl 0.7, reaction 0.6, termin.label 0.6, product 0.5, probe.surfac 0.5, bacteriorhodopsin 0.5, ladder 0.5, label.reaction 0.5, modif 0.4, bind 0.4, maldi.probe 0.4, site 0.4, ra 0.4, prepar 0.4, modif.mass 0.4, prepar.sampl 0.3)

(11) Cluster 9 (mirror 7.3, field 3.1, focus 1.8, ion.mirror 1.3, electr 1.3, extract 1.2, ion 1.2, delai 1.0, quadrat 0.9, reflect 0.7, energi 0.6, electr.sector 0.6, acceler 0.5, deficit 0.5, resolut 0.5, extract.field 0.5, sector 0.4, fring.field 0.4, fring 0.4, delai.extract 0.4)

(10) Cluster 10 (reflectron 2.7, tandem 2.4, projectil 1.9, ion 1.0, spectromet 0.9, fragment 0.8, mass.spectromet 0.8, cid 0.8, tandem.reflectron 0.8, reflectron.time.flight 0.8, reflectron.time 0.8, tryptophan 0.7, pth 0.6, flight.mass.spectromet 0.6, tandem.time 0.5, tandem.time.flight 0.5, spectromet.laser 0.5, mass.spectromet.laser 0.5, pattern 0.5, photodissoci 0.5)

(10) Cluster 11 (pom 2.2, oligom 1.8, end 1.6, end.group 1.5, polym 1.4, group 1.0, sampl.target.prepar 0.9, target.prepar 0.9, distribut 0.9, repeat 0.8, chlorin 0.7, sampl.target 0.7, methylmethacryl 0.7, antioxid 0.6, peak.area 0.6, poli 0.5, unit 0.5, repeat.unit 0.5, group.distribut 0.5, end.group.distribut 0.5)

(11) Cluster 12 (esi 5.0, polym 4.4, ftm 1.1, electrosprai.ioniz.esi 0.8, ioniz.esi 0.8, electrosprai.ioniz 0.7, ioniz.maldi.electrosprai 0.7, electrosprai 0.7, fluorin 0.7, maldi.electrosprai.ioniz 0.6, maldi.electrosprai 0.6, cyclic 0.5, mass.measur 0.5, measur.accuraci 0.5, mass.measur.accuraci 0.5, peak.correspond 0.4, high.mass 0.3, techniqu 0.3, discrimin 0.3, high 0.3)

(12) Cluster 13 (maldi.tof 3.1, 119, ESim: 0.008tof 2.0, maldi.tof.mass 1.8, weight 1.7, molecular.weight 1.6, distribut 1.4, tof.mass 1.3, polym 1.2, polydispers 1.2, tof.mass.spectrometri 0.9, flight.maldi 0.9, flight.maldi.tof 0.9, time.flight.maldi 0.9, averag.molecular.weight 0.8, molecular.weight.distribut 0.8, weight.distribut 0.8, poli 0.7, molecular 0.7, averag.molecular 0.7, averag 0.6)

(11) Cluster 14 (protein 6.9, databas 1.7, sequenc 1.0, gel 1.0, search 0.9, protein.sequenc 0.7, peptid 0.7, protein.mass.spectrometri 0.7, map 0.7, dimension 0.6, separ 0.6, new 0.6, peptid.mass 0.6, two.dimension 0.6, peptid.map 0.5, separ.protein.mixtur 0.5, separ.protein 0.5, electroblot 0.5, protein.mass 0.4, discrib 0.4)

(11) Cluster 15 (particl 4.0, saldi 2.8, tlc 2.7, suspens 1.2, analyt 0.9, surfac 0.9, carbon 0.8, powder 0.8, glycerol 0.7, graphit 0.7, plate 0.6, metal 0.6, spectra 0.5, tetracyclin 0.5, tlc.plate 0.5, mass.spectra 0.4, macromolecul 0.4, visibl 0.4, pore 0.4, low 0.4)

(10) Cluster 16 (puls 2.4, flight.mass.spectromet 1.4, spectromet.tof 1.2, mass.spectromet.tof 1.2, short 1.0, mass.spectromet 0.9, spectromet 0.9, new.gener 0.8, new.gener.time 0.8, tof.matrix.assist 0.8, paper.new.gener 0.8, spectromet.tof.matrix 0.8, gener.time.flight 0.8, tof.matrix 0.8, light.puls 0.7, light 0.7, paper.new 0.6, gener.time 0.6, aerosol 0.6, lam 0.5)

(10) Cluster 17 (polym 2.0, molecular.mass 1.0, maldi 0.8, wax 0.7, mean 0.7, mass.distribut 0.6, develop 0.6, copper.chlorid 0.6, distribut 0.6, chlorid 0.6, sec 0.6, copper 0.5, determin.molecular 0.5, molecular.mass.distribut 0.5, polystyren 0.5, sfc 0.4, telechel.polyisobutylen 0.4, classic 0.4, tof 0.4, chromatographi 0.4)

(11) Cluster 18 (ftm 1.9, fourier.transform.mass 1.4, transform.mass 1.4, resolut 1.1, fourier.transform 1.0, fourier 1.0, mass.resolut 1.0, 000 1.0, transform 0.9, digit.convert 0.9, dendrim 0.9, transform.mass.spectromet 0.8, digit 0.8, photodissoci 0.7, convert 0.6, ion 0.6, charg 0.4, deceler 0.4, lower 0.4, high.mass 0.4)

(12) Cluster 19 (oligom 7.9, glycol 3.1, ethylen 2.6, ethylen.glycol 2.1, ag 1.0, terephthal 0.9, kcal.mol 0.9, oligom.distribut 0.9, monom 0.8, kcal 0.7, poli 0.6, unit 0.6, varnish 0.5, poli.ethylen 0.5, bind.energi 0.5, ethylen.glycol.oligom 0.4, glycol.oligom 0.4, oligom.complex 0.4, complex 0.4, chain 0.4)

(10) Cluster 20 (cool 1.2, guest 0.8, intern 0.7, energi 0.7, molecul 0.7, ldi 0.7, expans 0.6, tryptophan 0.5, matrix.assist.desorpt 0.5, local 0.5, tea.co2 0.5, desorb 0.5, intern.energi 0.4, co2 0.4, assist.desorpt 0.4, molecular.cool 0.4, interfac 0.4, local.ioniz 0.4, veloc 0.4, temperatur 0.4)

(12) Cluster 21 (copolym 4.1, end 2.3, polyest 1.7, oligom 1.5, lactic.acid 1.3, lactic 1.3, nmr 1.1, pcl 1.0, group 1.0, reaction 0.9, end.group 0.9, microwav 0.8, chain.end 0.8, chain 0.7, block 0.7, unit 0.7, cyclic 0.6, carboxyl 0.6, propylen 0.6, ether 0.5)

(10) Cluster 22 (assai 1.0, techniqu 1.0, inform 0.5, techniqu.chemic 0.5, agent 0.5, mid.spectrum 0.5, mid 0.5, applic 0.4, method 0.4, structur 0.4, biopolym 0.4, gel.slice 0.4, slice 0.4, multipl 0.4, sampl 0.3, cell 0.3, agaros 0.3, chemic 0.3, gel 0.3, biolog 0.3)

(11) Cluster 23 (ammonium 2.3, acet 1.9, fragment 1.5, dhba 1.3, psd 1.2, ammonium.acet 1.2, oligosaccharid 1.2, maldi.tofm 1.0, oligonucleotid 1.0, tofm 0.9, acid 0.8, rutin 0.6, oligodeoxynucleotid 0.5, chca 0.5, cat 0.5, buffer 0.4, post.sourc 0.4, damgo 0.4, psd.fragment 0.4, structur.oligosaccharid 0.4)

(10) Cluster 24 (transferrin 2.4, arsen 1.8, seldi 1.4, pah 1.1, enzym 0.9, fibrillari 0.9, seldi.tof 0.8, sampl.prepar 0.7, prepar 0.7, magnet 0.5, doubl.bond 0.5, compound 0.4, spectrometri.maldi 0.4, mass.spectrometri.maldi 0.4, giant 0.4, insolubl 0.4, fibrillari.polym 0.4, tof 0.4, biomark 0.3, solvent 0.3)

(13) Cluster 25 (trap 9.8, ion 1.9, ion.trap 1.8, sourc 1.0, icr 0.8, axial 0.8, trap.mass 0.5, excit 0.5, ion.cyclotron.reson 0.5, cyclotron.reson 0.5, cyclotron 0.4, ion.cyclotron 0.4, reson 0.4, quadrupolar 0.4, quadrupolar.excit 0.4, ion.sourc 0.4, trap.potenti 0.4, 000 0.4, fticr 0.4, cool 0.4)

(13) Cluster 26 (acid 3.1, hydroxycinnam 1.4, hydroxycinnam.acid 1.1, ion.mode 1.1, 9aa 1.1, alpha.cyano 1.1, cyano 1.0, alpha 0.9, mode 0.8, alpha.cyano.hydroxycinnam 0.8, cyano.hydroxycinnam 0.8, cyano.hydroxycinnam.acid 0.8, dhb 0.8, magnet.sector 0.7, ctab 0.7, gangliosid 0.6, matric 0.6, sector 0.6, methyl 0.5, mbt 0.5)

(12) Cluster 27 (eject 4.2, sputter 1.3, recent 1.0, mechan 0.9, solid 0.7, surfac 0.7, mechan.laser 0.7, desorpt.method 0.6, spectrometri.particular 0.6, mass.spectrometri.particular 0.6, desorpt.ion 0.5, method 0.5, practic 0.5, thermal 0.5, laser.induc 0.4, induc 0.4, applic 0.4, experiment 0.4, yield 0.3, particular 0.3)

(12) Cluster 28 (ion 1.0, formazan 0.9, matric 0.9, analyt 0.9, phase 0.7, pre 0.7, state 0.6, process 0.6, pre.form 0.6, atom 0.6, analyt.ion 0.5, photoion 0.5, form.ion 0.5, atom.substitut 0.5, heavi.atom 0.5, maldi 0.5, paa 0.4, ion.format 0.4, ablat 0.4, ioniz.process 0.4)

(13) Cluster 29 (sampl 2.0, fluenc 1.7, protein 1.3, analyt 1.2, nitrocellulos 0.7, multimer 0.7, molecular.ion 0.6, matrix.analyt 0.6, laser.fluenc 0.6, ion 0.6, peak 0.6, irradi 0.6, exposur 0.5, sampl.exposur 0.5, laser.irradi 0.5, microscopi 0.4, resolut 0.4, mass.extend 0.4, coumarin 0.4, analyt.molecul 0.4)

(12) Cluster 30 (substrat 3.3, standard 2.1, dna 1.2, nafion 1.1, nafion.substrat 1.1, film 0.9, rsd 0.8, intern.standard 0.8, segment 0.7, cerevisia 0.6, chlormequat 0.5, thin 0.5, csa 0.4, dna.segment 0.4, ablat 0.4, intern 0.4, peak 0.4, pac 0.3, silver 0.3, laser.ablat

0.3)

(13) Cluster 31 (liquid 2.4, absorb 2.0, liquid.matric 1.0, glycerol 1.0, matric 0.9, matrix.solut 0.9, sampl 0.8, solid 0.8, absorb.concentr 0.7, solut 0.7, absorpt 0.6, matrix.system 0.6, coat.protein 0.5, reproduc 0.5, liquid.matrix 0.5, support 0.4, protein 0.4, solid.matrix 0.4, liquid.support 0.4, shot 0.4)

The CLUTO algorithm then aggregates the clusters in a hierarchical taxonomy. Overall, Level 1, the total database (ionization, MALDI, protein, peptide, polymer) contains 344 records, with a broad focus of MALDI, bio-molecular, and non-biomolecular applications. Level 2 contains the first major categorical split of two categories: Applications and Analytical Process. There are 131 records in Applications (oligomer, protein, polymer, peptide, molecular weight, MALDI, TOF), focused on large bio-molecules, oligomers (small polymers), and large polymers. Additionally, there are 213 records in Analytical Process (ionization, analyte, sample, MALDI, resolution, fragment, acid, matrix), focusing on the charging process and the sample preparation.

Level 3 contains the next categorical split of 4 categories: Bio-molecules, Non-bio-molecules, Sample Preparation, and Mass Resolution. The Applications category sub-divides evenly into Bio-molecules and Non-bio-molecules. There are 66 records in Bio-molecules (protein, peptide, amino acid sequence), focused on proteins, peptides, and amino acid sequencing. There are 65 records in Non-bio-molecules (oligomer, polymer, molecular weight), focused on oligomers and polymers. This Non-bio-molecules category does not appear in the Fenn citing papers, at least as a dominant theme.

The Analytical Process category sub-divides into Sample Preparation and Mass Resolution. There are 95 records in Sample Preparation (analyte, matrix, acid, ion, proton), focused on the steps leading to ionization, especially on preparation of the matrix. There are 118 records in Mass Resolution (ionization, resolution, trap, energy, mass spectrometer), focused on the control of mass spectrometer fields and energies necessary to increase the precision of mass determination.

4. SUMMARY

Publication Bibliometrics

There were 1628 papers that cited Fenn's 1989 paper, and 410 papers that cited Tanaka's 1988 paper. Because the SCI did not start to publish Abstracts until 1991,

and because not all citing papers have Abstracts, only 1433 of the Fenn citing papers in the SCI database contain Abstracts, and only 344 of the Tanaka citing papers contain Abstracts. The bibliometrics analyses are performed on the total number of citing papers, whereas the computational linguistics are performed on those papers with Abstracts.

Author Frequency Results

For the Fenn citing papers, seventeen of the 22 most prolific authors are from the USA, two are from Australia, two are from Denmark, and one is from Japan. Fifteen are from universities, three are from research institutes, and four are from industry.

For the Tanaka citing papers, eight of the 23 most prolific authors are from the USA, and the remainder are from Europe, mainly central Europe. Twenty are from universities, and three are from research institutes. No authors are common to the two lists of prolific citing authors. Why are there no prolific citing authors from Japan, and why are there no prolific citing authors from industry, for Tanaka's research? This is surprising, since Tanaka is both from Japan and industry.

Journal Frequency Results

For both the Fenn and Tanaka citing papers, the most prolific journals focus on mass spectrometry, chemistry, and biology. Three journals stand out as the first tier for containing the most citing papers: ANALYTICAL CHEMISTRY, JOURNAL OF THE AMERICAN SOCIETY FOR MASS SPECTROMETRY, RAPID COMMUNICATIONS IN MASS SPECTROMETRY. Twelve journals are in common between the two lists. The prolific Fenn citing journals not in common tend to focus on biology/ biochemistry (ANALYTICAL BIOCHEMISTRY, BIOCHEMISTRY, PROTEIN SCIENCE, EUROPEAN JOURNAL OF BIOCHEMISTRY), while the prolific Tanaka citing journals not in common tend to focus on the technique/ instrumentation (REVIEW OF SCIENTIFIC INSTRUMENTS, ORGANIC MASS SPECTROMETRY, EUROPEAN MASS SPECTROMETRY). This observation supports the later document clustering finding of the greater emphasis on bio-molecules in the Fenn citing papers relative to the Tanaka citing papers.

Institution Frequency Results

Of the twenty institutions producing the most Fenn citing papers, seventeen are from North America, one from Europe, and two from the Far East. Seventeen are

universities, and three are research institutes. Of the twenty institutions producing the most Tanaka citing papers, twelve are from the USA, seven are from Europe, and one is from Japan. Eighteen are universities, one is a research institute, and one is from industry. Four institutions are in common between the two lists: UNIV CAL SAN FRANCISCO, INDIANA UNIV, ROCKEFELLER UNIV, OSAKA UNIV.

Country Frequency Results

The USA clearly dominates in country output. The next tier is high on both Fenn and Tanaka citing lists (GERMANY, ENGLAND, JAPAN, CANADA), with Switzerland appearing high on the Tanaka citing list. Thus, while Japan is not very visible in terms of prolific citing authors or institutions, especially with respect to Tanaka's paper, it has reasonable representation in terms of country citations. This implies a diverse group of citing authors in Japan, with the exception of the group at Osaka University.

In terms of absolute numbers of co-authored Fenn-citing papers, the USA major partners are Canada, Japan, Germany, England, and France. Additionally, the USA is the major partner for ten of the countries, the exceptions being Australia, Belgium, Holland, and China.

In terms of absolute numbers of co-authored Tanaka-citing papers, the USA major partners are Germany, Canada, England, and Japan. Additionally, the USA is the major partner for nine of the countries, the exceptions being Australia, Austria, Holland, Scotland, and Switzerland.

Citation Statistics on Authors, Papers, and Journals

The second group of metrics presented is counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics, *much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers*.

Author Citation Frequency Results

In the Fenn citing papers, Fenn is cited almost twice as much as the next ranked author. This is due to the citation of Fenn's other related papers between 1984 and 1989, in addition to the citation of the Science article. The next tier, RD Smith and JA Loo, was a very prolific and highly cited group working on different mass spectrometry techniques, including electrospray ionization.

In the Tanaka citing papers, Tanaka actually ranks third in number of first-author citations. M. Karas of Frankfurt ranks first (along with F. Hillenkamp of Muenster, who co-authored many of these papers with Karas). This is due to three factors. First, in 1985, Karas, in conjunction with Hillenkamp, showed that a “strongly absorbing matrix at a fixed laser wavelength” could be used to vaporize small molecules without chemical degradation. Second, in 1988, Karas and Hillenkamp reported a MALDI approach applied to proteins shortly after Tanaka’s paper was published. Thus, the papers that cite Tanaka’s paper also tend to cite the groundwork papers of Karas/ Hillenkamp as well as their large molecule mass determination papers. Third, Karas and Hillenkamp were in the top tier of Tanaka citing authors, as well as prolific in their own right relative to Tanaka, and had more opportunity to cite their own foundational work in the papers in which they also cited Tanaka. Additionally, due to a series of highly-cited papers by RC Beavis (along with his co-author B. Chait) in the early 1990s on laser desorption mass spectrometry, many of the papers that cite Tanaka tend to multiply cite Beavis/ Chait. This large co-citation of Karas/ Hillenkamp and Beavis/ Chait with Tanaka was mentioned in the Background. It was shown that, of the top fifty cited laser desorption mass spectrometry papers produced in the early high growth years, Tanaka’s paper was referenced in fifteen, while the Beavis/ Chait papers were referenced in 37 and the Karas/ Hillenkamp papers were referenced in 38.

There are five names in common between the two lists of most highly cited authors in the Fenn and Tanaka citing papers (FENN, SMITH, KARAS, BEAVIS, HILLENKAMP). All five have made broad contributions to mass spectrometry.

Of the 21 most cited authors in the Fenn citing papers, fourteen are from universities, three are from research institutions, and four are from industry. Of the 21 most cited authors in the Tanaka citing papers, sixteen are from universities, one is from a research institute, and four are from industry. This relatively high fraction (~20%) of cited papers from industry suggests relatively applied citing papers. The validity of this assumption is confirmed in the sections on temporal citing patterns and document clustering.

Finally, while Central Europe plays a modest role in the reference source for the Fenn list, it continues to play a much stronger role for the Tanaka list.

The citation data for authors and journals represents citations generated only by the specific records extracted from the SCI database for this study. It does not represent all the citations received by the references in those records; these references in the

database records could have been cited additionally by papers in other technical disciplines.

Document Citation Frequency Results

For the twenty most cited documents in the Fenn citing papers, Analytical Chemistry contains the most highly cited documents (six). For the twenty most cited documents in the Tanaka citing papers, both Analytical Chemistry and Rapid Communications in Mass Spectrometry each contain five of the most highly cited documents.

All of the journals containing these most highly cited documents are fundamental science journals, and most of the topics have a fundamental science theme. Of the most highly cited documents in the Fenn citing papers, nine are from the 80s, eight are from the 90s, and one each from the 70s and 60s. Of the most highly cited documents in the Tanaka citing papers, twelve are from the 90s, seven are from the eighties, and one is from the 50s. These numbers reflect dynamically evolving disciplines, with many of the seminal works coming from recent times.

From the lists of references in the Fenn citing papers, about thirty percent of the papers address the phenomena underlying electrospray (ION SOURCE-FREE JET, ELECTROSPRAY INTERFACE, MULTIPLY-CHARGED IONS, MACROION BEAMS, CHARGED DROPLET ION EVAPORATION), about twenty five percent address the electrospray technique (ELECTROSPRAY IONIZATION, HYBRID MASS SPECTROMETRY), about thirty percent address applications (LARGE POLYPEPTIDES, PROTEINS, RECEPTOR LIGAND COMPLEXES), and a few address laser desorption. From the lists of references in the Tanaka citing papers, about fifteen percent of the papers address the laser desorption approach and associated phenomena, about ten percent address the electrospray technique, and the remainder address applications (LARGE PROTEINS, NONVOLATILE COMPOUNDS, BIOPOLYMERS, LARGE BIOMOLECULES, SYNTHETIC POLYMERS), mainly using the MALDI technique. The relatively large numbers of cited papers related to applications are consistent with the observation in the previous section that a relatively substantial number of highly cited authors were from industrial organizations.

Journal Citation Frequency Results

Sixteen of the top twenty most highly cited journals are in common between the two lists. Those not in common from the list of most cited journals in Fenn citing papers are: ELECTROPHORESIS, NATURE, METHODS ENZYMOLOGY, JOURNAL OF

CHROMATOGRAPHY A. Those not in common from the list of most cited journals in Tanaka citing papers are: BIOMEDICAL ENVIRONMENTAL MASS, MACROMOLECULES, CHEM PHYS LETTERS, BIOLOGICAL MASS SPECTROMETRY.

The list of journals containing the most Fenn citing papers, and the list of most cited journals in the Fenn citing papers, had thirteen journals in common. The list of journals containing the most Tanaka citing papers, and the list of most cited journals in the Tanaka citing papers, also had thirteen journals in common.

Temporal Citing Patterns

In the original citation mining paper, two characteristics of the citing papers were evaluated as a function of time. These were: 1) the level of development of the work reported in the citing paper (basic research, applied research, technology development) and 2) the alignment between the technical thrusts of the citing paper and the cited paper (strongly aligned, partially aligned, not aligned). These temporal results provided useful insights to the evolution of the nature of the citing papers as time proceeded, and it was decided to perform a similar analysis for the present paper. In order to have sufficient data to evaluate these two characteristics credibly, only those citing papers with Abstracts were included in the analysis (1433 citing papers for Fenn, 344 citing papers for Tanaka)

A two character metric was used to quantify the above two characteristics. The first character represented level of development, and ranged from one (most fundamental research) to three (applied technology development). The second character represented degree of alignment, and ranged from one (fully aligned) to three (non-aligned). Table 2 presents the temporal citing results. Table 2A presents the results for the Fenn citing papers (Table 2A-1-normalized), and Table 2B presents the results for the Tanaka citing papers (Table 2B-1-normalized). The first column in each table is the two character metric. The matrix elements Mij represent the number of citing papers with metric i published in year j. For example, in Table 2A, there were 25 citing papers in 2001 that were both fundamental research and fully aligned with the theme of the (Fenn) cited paper.

TABLE 2A – FENN CITING PAPER CHARACTERISTICS VS TIME

METRIC	ALL YEARS	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	451	18	25	30	49	52	68	69	27	36	28	23	24	2

12	328	16	13	12	11	29	44	40	60	44	40	10	8	1
13	121	6	12	9	5	9	10	6	10	21	14	14	5	
21	299	13	21	24	43	42	30	18	21	23	22	19	22	1
22	170	18	38	56	14	10	8	6	7	8	3	1	1	
23	49	16	14	9	3	2	4	1						
31	10	1	2	5	1	1								
32	4		3	1										
33	1		1											
TOTAL->	1433	88	129	146	126	145	164	140	125	132	107	67	60	4

TABLE 2A-1 – FENN CITING PAPER CHARACTERISTICS VS TIME (NORM)

METRIC	ALL YEARS	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	0.31	0.2	0.19	0.21	0.39	0.36	0.41	0.49	0.22	0.27	0.26	0.34	0.4	0.5
12	0.23	0.18	0.1	0.08	0.09	0.2	0.27	0.29	0.48	0.33	0.37	0.15	0.13	0.25
13	0.08	0.07	0.09	0.06	0.04	0.06	0.06	0.04	0.08	0.16	0.13	0.21	0.08	0
21	0.21	0.15	0.16	0.16	0.34	0.29	0.18	0.13	0.17	0.17	0.21	0.28	0.37	0.25
22	0.12	0.2	0.29	0.38	0.11	0.07	0.05	0.04	0.06	0.06	0.03	0.01	0.02	0
23	0.03	0.18	0.11	0.06	0.02	0.01	0.02	0.01	0	0	0	0	0	0
31	0.01	0.01	0.02	0.03	0.01	0.01	0	0	0	0	0	0	0	0
32	0	0	0.02	0.01	0	0	0	0	0	0	0	0	0	0
33	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0

TABLE 2B – TANAKA CITING PAPER CHARACTERISTICS VS TIME

METRIC	ALL YEARS	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	136	9	9	12	6	15	18	15	11	11	14	6	10	0
12	86	9	10	4	10	7	9	9	10	5	6	6	1	
13	50	2	2	5	4	7	10	6	2	5	4	1	2	
21	43	1	1	4	3	4	1	6	5	5	5	3	5	
22	20	1	4	4		3		2	1	2	2	1		
23	6	1	2	1		2								
31	2			2										
32	0													
33	1		1											
TOTAL->	344	23	29	32	23	38	38	38	29	28	31	17	18	0

TABLE 2B-1 – TANAKA CITING PAPER CHARACTERISTICS VS TIME (NORM)

METRIC	ALL YEARS	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	0.4	0.39	0.31	0.38	0.26	0.39	0.47	0.39	0.38	0.39	0.45	0.35	0.56	
12	0.25	0.39	0.34	0.13	0.43	0.18	0.24	0.24	0.34	0.18	0.19	0.35	0.06	
13	0.15	0.09	0.07	0.16	0.17	0.18	0.26	0.16	0.07	0.18	0.13	0.06	0.11	
21	0.13	0.04	0.03	0.13	0.13	0.11	0.03	0.16	0.17	0.18	0.16	0.18	0.28	
22	0.06	0.04	0.14	0.13	0	0.08	0	0.05	0.03	0.07	0.06	0.06	0	

23	0.02	0.04	0.07	0.03	0	0.05	0	0	0	0	0	0	0
31	0.01	0	0	0.06	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0.03	0	0	0	0	0	0	0	0	0	0

In aggregate, the Tanaka citing papers have a moderately greater concentration in basic research (first metric character of unity) than the Fenn citing papers, 0.80 normalized vs. 0.62 normalized. The Tanaka citing papers have a greater concentration in the most non-aligned category (second metric character of three) than the Fenn citing papers, 0.17 normalized vs. 0.11 normalized. The Fenn citing papers have a greater concentration in the applied research most-aligned category (metric of 21) than the Tanaka citing papers, 0.21 vs. 0.13 normalized. These three findings corroborate the most prolific authors bibliometrics results, which showed almost twenty percent of the most prolific Fenn citing authors were from industry, whereas none of the most prolific Tanaka citing authors were from industry.

The temporal evolution shows that about a decade is required before the applied technology citing papers become evident. It should be stressed that these are the directly citing technology papers, i.e., papers that cited the original Fenn or Tanaka papers. It is possible that indirectly citing technology papers (i.e., papers that did not cite Fenn or Tanaka's original paper, but rather cited other papers that had cited the Fenn or Tanaka original papers) appeared earlier, but this higher generation bibliometric analysis was beyond the scope of the present study.

One other citation mining study has been performed. Emphasized in that study, and comparable in spirit to the present study, was a detailed analysis of the 1992 Science paper of Jaeger and Nagel on dynamic granular systems. That paper was a very fundamental research paper focused on the basic physics of flowing granular systems. The normalized temporal evolution of the citing papers of that study is shown in Table 3. Relative to the Fenn and Tanaka citing papers, the Jaeger and Nagel citing papers have a substantially higher basic research fraction in aggregate. There was a four year lag time before any applied citing papers emerged. Beyond what the numbers portray, the Jaeger and Nagel citing papers reached a wider variety of more extreme non-aligned categories than the Fenn or Tanaka citing papers (e.g., earthquakes, avalanches, traffic congestion, war games, flow immunosensors, shock waves, nanolubrication, thin film ordering). Chi-tests confirmed the validity of the differences between the Fenn-Tanaka citing papers and the Jaeger and Nagel citing papers, and between the Fenn and Tanaka citing papers as well.

TABLE 3 – JAEGER AND NAGEL CITING PAPER CHARACTERISTICS VS TIME (NORM)

METRIC	ALL YEARS	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
11	0.78	0.67	0.69	0.75	0.75	0.84	0.77	0.85	0.85	0.75		
12	0.15	0.17	0.21	0.2	0.18	0.08	0.17	0.09	0.07	0		
13	0.04	0	0.02	0	0.05	0.04	0.06	0.06	0.07	0.25		
21	0.01	0	0	0.03	0.02	0	0	0	0	0		
22	0.01	0	0.06	0	0	0.02	0	0	0	0		
23	0.01	0.17	0.02	0.03	0	0	0	0	0	0		
31	0	0	0	0	0	0	0	0	0	0		
32	0	0	0	0	0	0.02	0	0	0	0		
33	0	0	0	0	0	0	0	0	0	0		

Computational Linguistics (Taxonomy Generation)

Three statistically-based clustering methods, factor matrix, multi-link aggregation, and partitional document clustering, were used to develop taxonomies. They each offered a modestly different perspective on taxonomy category structure. Neither of the three approaches is inherently superior, and all should be viewed as complementary.

For both the Fenn and Tanaka citing paper databases, the words contained in the citing paper Abstracts were extracted by the Vantage Point software, and sorted by frequency of occurrence. The highest frequency high technical content words were identified by inspection. Very similar words were consolidated (e.g., singulars/ plurals, full spellings/ acronyms, very strong synonyms).

<u>Factor Matrix Clustering</u>

A correlation matrix of the 253 resultant consolidated words was generated, and a factor analysis was performed using the WINSTAT statistical package (an Excel add-in). The eigenvalue floor was set equal to unity to insure that each resulting factor provide value-added information, and a factor matrix consisting of 42 factors (columns) and 253 words (rows) resulted.

Each matrix element Mij is known as the factor loading, and is a measure of the contribution of word i to factor j. A factor represents a technical theme, and some combination of the factors represents a taxonomy. There are cases where words have high loadings in multiple factors (e.g., RECOMBINANT has a value of .46 in factor 1, .37 in factor 10, and .21 in factor 37), and they usually (not always) tend to be situated in the factor where they have the highest loading. These high loading multi-factor

words do, however, serve as a link among the factors, and cause the factors to overlap.

Overall, the factor matrix required more factors, and had more overlap among factors, than in previous text mining studies. These previous studies focused on papers related to a focused theme, not to a cited paper as in the present study. The citing paper database is more diverse and fragmented, since it incorporates many different types of applications. Its component papers tend to mix both application and technique/ technology development, as opposed to the much stronger focus on technique/ technology development that characterized the previous studies. This added diversity, and the mixing of technology development with applications, translates into a larger number of factors that have numerous overlaps.

The factors in the matrix are ordered by cohesiveness. Factor 1 is larger in extent and more focused than the other factors. As the factor numbers increase, the factors contain less words and their theme becomes more diffuse.

Multi-Link Clustering

A symmetrical co-occurrence matrix of the 253 highest frequency high technical content words was generated. The matrix elements were normalized using the Equivalence Index ($E_{ij}=C_{ij}^2/C_i*C_j$, where C_i is the total occurrence frequency of the ith phrase, and C_j is the total occurrence frequency of the jth phrase, for the matrix element ij), and a multi-link clustering analysis was performed using the WINSTAT statistical package. The Average Linkage method was used. The hierarchical structure of the taxonomy is guided by the branching structure of the dendrogram, a tree-like structure output by the software. As the dendrogram progresses from one hierarchical level to the next downward level, each branch divides into two parts. Thus, the highest level of the taxonomy consists of two clusters, the next level consists of four clusters, and so on.

Fenn Citing Papers Multi-Link Taxonomy

Based on integrating the results from the factor matrix and multi-link clustering, the Fenn citing papers could be categorized into the following reasonable taxonomy, with the thrust areas delineated.

Separation
[chromotography, solvents]

- separation techniques used to purify and separate digested macromolecules, proteins or polypeptides prior to introduction into, and identification by, various mass spectrometry methods. For biochemical materials, two separation techniques are generally used; high pressure liquid chromatography and gel electrophoresis.

- role of solvent compositions in the liquid and gel chromatographic separation process.

Ionization Source/ Processes
[CID, droplet charge, alkali metal cations, coulomb repulsion, proton transfer, labile proton exchange]

- analysis of species generated from proteins and other macromolecules by collision-induced dissociation in quadrupole ion trap coupled tandem mass spectrometry. Ion mass spectra analysis of the resulting fragments determines the structure (typically) of metabolites, peptides and polypeptides and affords information on the reconstruction of parent molecules or proteins.

- effect of electrostatic and liquid properties, and flow variables, on the size and charge of droplets ejected in conical sprays from capillaries prior to solvent elimination and injection of the digested substrate material fragments into the mass spectrometer for characterization.

- use of negative ion mass spectrometry to study the structure of sodium and other alkali metal salts, with emphasis on adduction of alkali metal compounds with anions of the larger alkali metal ions.

- influence of coulomb repulsion on the reaction and dissociation rates of singly and doubly charged ions (including protonated and deprotonated ions) within the mass spectrometry system.

- proton transfer reactivity in the gas-phase, and reaction of singly- and multiply-protonated molecules.

- determining exchange numbers and rates of labile protons (hydrogen/ deuterium exchange) and ligands by both electrospray ionization mass spectrometry and NMR.

Mass Analyzer

[ion cyclotron resonance, atmospheric pressure sources, magnetic sector instruments, quadrupole ion trap]

- Fourier transform ion cyclotron resonance mass spectrometry, including different ion source configurations and cyclotron resonance excitation, for instruments of higher mass resolution.

- collision-activated dissociation at the vacuum/ atmospheric pressure interface of liquid chromotography mass spectrometry, especially systems with atmospheric pressure ionization sources.

- mass spectrometric systems efficiency, based on magnetic sector instrumentation, for large protein mass and structure determination.

- storage and accumulation of large source biopolymer ions in a quadrupole ion trap, and subsequent injection of these ions into the flight tube of a time-of-flight mass spectrometer. This process converts the typically continuous source ion beam into a higher density pulsed ion beam for the mass spectrometer, resulting in higher resolution and sensitivity.

Mass Spectrometer System
[MALDI, electrospray IMS, internal standards]

- matrix-assisted laser desorption ionization time-of-flight mass spectrometry for peptide mass fingerprinting, followed by post source decay analysis for more detailed characterization of amino acid sequences.

- use of electrospray ionization mass spectrometry for the detection of mass spectra of large molecules in solution.

- use of ionization mass spectrometry for the quantitative determination of organic compounds in plasma, making use of internal standards for quantification or standardization

Applications Predominantly Biomedical
[amino acid structure, conformation, E Coli gene expression, generic proteins]

- protein construction, characterization and structure through analysis of digested and recombined amino acid, peptide and polypeptide sequences and residues. General

disulfide cleavage, trypsin digestion, and determination of C- and N- terminal groups, afford methods for reconstructive mapping of fragments and segments of proteins and other macromolecules.

- interaction between protein digestion (ubiquitin, lysozyme) and aspects of the structure/ conformation determination process using ion mobility mass spectrometry.

- genetically expressed proteins in Escherichia Coli cells and recombination technology.

- fundamental generic sources of protein types and functions used in the determination of structural properties.

Applications Other, including some Biomedical
[noncovalent complexes, carbohydrate structures, analyte solution mass spectra]

- use of soft ionization mass spectrometry for studying noncovalently bound complexes, including interaction strength. Emphasis is on deduction of the stoichiometry of the binding partners from the molecular weight measurement, and use of the mass spectrometry-based method to assess the affinity of such interactions. Focuses on the non-ionic protein and macromolecule conformational and structural interactions.

- defining the structural heterogeneity of the carbohydrate oligosaccharide moiety of glycoprotein, using ionization mass spectrometry.

- mass spectra of aqueous analyte solutions with varying concentrations of ammonium acetate.

<u>Tanaka Citing Papers Multi-Link Taxonomy</u>
Based on integrating the results from the factor matrix and multi-link clustering, the Tanaka citing papers could be categorized into the following reasonable taxonomy, with the thrust areas delineated.

Separation
[capillary/ gel chromatography]

- combination of capillary and gel permeation chromatography for the separation of materials for subsequent injection into the mass spectrometer.

Sample Preparation
[alpha-cyano-4-hydroxycinnamic matrices, thin film substrates, carbon/ metallic matrices, aqueous matrix solutions, dithranol matrices/ silver salts dopants]

- identifying the spectra complexity and signal strength differences in positive-ion and negative-ion modes. Emphasis is on use of the matrices alpha-cyano-4-hydroxycinnamic acid and 2,5-dihydroxybenzoic acid, including analysis of the strong dependence of fragmentation on the nature of the matrix and on the presence or absence of water in the matrix solvent.

- deposition of organic samples on stable thin precious metal film substrates, to minimize background interference with analyte ion peaks, facilitate deposition of samples from a variety of solvent systems, and generate analyte adduct ions in some cases. Emphasis is on desorption and ionization by Nd:YAG laser irradiation, and maximizing light absorption by film thickness variation.

- role of carbon and other metallic materials as a matrix target plate component, acting as an energy transfer agent by enhanced radiation absorption, and enhancing desorption of solvent and analyte ions.

- use of analyte and matrix aqueous solutions for MALDI, especially water and glycerol matrix additions, and the resultant ion signal stability for repeated laser shots.

- MALDI analysis of synthetic polymers, especially polystyrene, with use of dithranol as a matrix and silver salt ions as dopant to enhance the cationization of polystyrene through the formation of adduct complexes.

Ionization source/ process
[fast atom bombardment, CID, UV-infrared, multiply-charged ESI, protonated alkali adducts, gas-phase affinity, phase transition, analyte molecule ejection]

- use of fast atom bombardment mass spectrometry analysis.

- use of tandem mass spectrometry with collision-induced dissociation to provide structural information for unknown sample molecules.

- direct comparisons of the effectiveness of UV and IR lasers in MALDI mass spectrometry, especially YAG lasers, and concentrates on the role of water in the matrix-analyte as a major laser energy absorber.

- formation of singly and multiply charged molecular ions via the field-assisted ion evaporation mechanism during electrospray ionization.

- use of counter ions that accompany much of the digestion process used in peptides and macromolecules prior to introduction into an ion mass spectrometer.

- effects of proton affinity of MALDI matrices on the relative protonation of analytes from radical matrix molecular ions or protonated matrix ions.

- effect of solid-to-gas phase transition from laser irradiation on subsequent ionization and lift-off velocity of analyte molecules.

- modeling the ejection of analyte molecules at different laser fluences, concentrating on the dependence of the yield on fluence near ablation threshold to distinguish between ejection models.

Mass Analyzer
[ion cyclotron resonance, accelerating voltage focusing, reflecting mirror fields, signal processing FWHM, high S/N negative ion mode]

- laser desorption Fourier transform ion cyclotron resonance mass spectrometry, concentrating on cooling and axializing the MALDI-generated ions by azimuthal quadrupolar excitation in the presence of collisions with neutral atoms in the source compartment of a dual ion trap, followed by detection of the axialized ions at much lower pressure and much higher mass resolving power.

- increasing the resolving power of time-of-flight spectra, using delayed extraction of MALDI-generated ions and high accelerating voltage reflecting mirror fields to focus the ions and extend their flight paths.

- extending the flight path by the use of reflecting electrostatic mirrors to increase the time-of-flight of MALDI molecular ions for increased mass resolution, concentrating on delayed extraction from high acceleration voltage sources.

- combination of substrate materials (sinapinic acid, bovine) with the optical processes for MALDI analysis of large macromolecules, concentrating on the FWHM intensity and breadth of protonated peaks.

- solution degradation kinetics, emphasizing negative-ion mode operation with deprotonated molecular ions for high signal-to-noise mass spectra.

Mass Spectrometer System
[MALDI]

- using matrix-assisted laser desorption ionization mass spectrometry with time-of-flight mass analyzer for molecular weight determination.

Applications Predominately Biomedical
[amino acid sequencing]

- characterization of proteins' structures and properties through their component peptide structures by molecular weight determination and peptide amino acid sequences and residues, using MALDI mass spectrometry as a central technique. It includes peptide mapping of recombinant proteins to obtain structural and conformational information. Peptides are generated by protein digestion and typically separated by high pressure liquid chromatography.

Applications Other
[cyclic oligomer and synthetic polymer characterization, synthetic polymer characterization, covalent complex interactions]

- characterizing oligomer and polymer average molecular weights from MALDI, and comparing against NMR and GPC analysis. Emphasis is on cyclic oligomers, and end groups of synthetic polymers, especially on the relation between the composition of the terminal group on a polymer chain and the ion yields. Synthetic polymers such as polystyrene and poly (ethylene glycol) are emphasized

- characterizing average molecular weights of oligomers, and end groups of synthetic polymers using MALDI. Synthetic polymers such as polystyrene and poly (ethylene glycol) are emphasized.

- determining the interaction strength in non-covalently bound complexes, and the influence of hydrophobic interactions to establish differences between solution-phase and gas-phase binding energies.

Partitional Document Clustering

Document clustering is the grouping of similar documents into thematic categories. The approach presented here is based on a partitional clustering algorithm contained within a software package named CLUTO.

<u>Fenn Citing Papers Document Clustering Taxonomy</u>

The high value words extracted in each category are shown in parentheses. Overall, the main category (ionization, protein, peptide, charge, ESI, complex, sequence, acid), Level 1, contains 1431 records, with a broad focus of bio-molecular applications and the ionization-charge components of the mass detection and analysis process. Level 2 contains the first major categorical split of two categories: Applications and Ionization Process. There are 532 records in Applications (protein, peptide, sequence, MALDI, binding, DNA, acid, amino), focused on large bio-molecules. Additionally, there are 899 records in Ionization Process (ionization, charge, proton, solutions, electrospray ionization, state, fragment, dissociation), focusing on the charging process and charge state, as well as the sample solution prior to ionization.

Level 3 contains the next categorical split of 4 categories: Bio-molecule Structure, MALDI Protein Mapping, Ionization, and Sample Preparation. The Applications category sub-divides into Bio-molecule Structure and MALDI Protein Mapping. There are 349 records in Bio-molecule Structure (protein, peptide, binding, sequence, residue, beta, alpha, amino), focused on proteins, peptides, binding states, and amino acid sequencing.

There are 183 records in MALDI Protein Mapping (MALDI, protein, matrix, laser, desorption), focused on the use of MALDI for protein mapping. Sampling of these records shows the main focus to be MALDI, with Fenn/ ESI appearing mainly as a reference. Appearance of MALDI papers in the Fenn citing papers implies that either ESI is being cited as a MALDI alternative for Protein Mapping or that ESI is being cited historically as a demonstration that large bio-molecule mass measurements were possible.

Who are the MALDI researchers most cited in the Fenn citing papers? As the list of most cited authors shows, the main soft laser desorption researchers listed are Karas/

Hillenkamp. Tanaka does not appear in the top twenty list. To test whether this result applies beyond the Fenn citing papers, in a more recent context, a database of 300 papers was generated from the SCI. The query used was the same as in the Background (laser and desorption and (ion* or mass spectrometry)), and the records were the most recent prior to October 2002 (so as not to be influenced by the Nobel awards). After the elimination of (few) self-citations, the citation results were as follows: Karas-70 citations; Hillenkamp-25 citations; Tanaka-18 citations; Beavis-12 citations. 79% of the Karas citations were pre-1989 (1985-1988). These results mirror those using MALDI as the query term. Remembering that the SCI provides the first author in citation print-outs, and most of the early soft laser desorption papers of Karas and Hillenkamp were joint, it appears that the most referenced early works on soft laser desorption/ MALDI are those of Karas/ Hillenkamp. As shown in the Background, this was true over a decade ago, and as shown in this paragraph, it remains true today.

The Ionization Process category sub-divides into Ionization and Sample Preparation. There are 398 records in Ionization (ionization, charge, proton, charge state, dissociation, energy, fragment), focused on characteristics of the charged state. There are 501 records in Sample Preparation (droplet, solution, metal, ion, capillary, complex, liquid), focused on the process and components preparatory to ionization.

<u>Tanaka Citing Papers Document Clustering Taxonomy</u>

Overall, Level 1, the total database (ionization, MALDI, protein, peptide, polymer) contains 344 records, with a broad focus of MALDI, bio-molecular, and non-biomolecular applications. Level 2 contains the first major categorical split of two categories: Applications and Analytical Process. There are 131 records in Applications (oligomer, protein, polymer, peptide, molecular weight, MALDI, TOF), focused on large bio-molecules, oligomers (small polymers), and large polymers. Additionally, there are 213 records in Analytical Process (ionization, analyte, sample, MALDI, resolution, fragment, acid, matrix), focusing on the charging process and the sample preparation.

Level 3 contains the next categorical split of 4 categories: Bio-molecules, Non-bio-molecules, Sample Preparation, and Mass Resolution. The Applications category sub-divides evenly into Bio-molecules and Non-bio-molecules. There are 66 records in Bio-molecules (protein, peptide, amino acid sequence), focused on proteins, peptides, and amino acid sequencing. There are 65 records in Non-bio-molecules (oligomer, polymer, molecular weight), focused on oligomers and polymers. *This category does not appear in the Fenn citing papers, at least as a dominant theme*.

The Analytical Process category sub-divides into Sample Preparation and Mass

Resolution. There are 95 records in Sample Preparation (analyte, matrix, acid, ion, proton), focused on the steps leading to ionization, especially on preparation of the matrix. There are 118 records in Mass Resolution (ionization, resolution, trap, energy, mass spectrometer), focused on the control of mass spectrometer fields and energies necessary to increase the precision of mass determination.

5. CONCLUSIONS

Citation Mining can provide a comprehensive picture of the research citing community, including its technical infrastructure, technical thrusts and their relationships. If the citing community is assumed to represent the user community, then Citation Mining provides a reasonable picture of the user community. In this report, only the direct citations (first-order) were used to represent the user community. If indirect (second-order and higher) citations are included as well (i.e., the papers that cite the papers that cite the original paper, and their parent/ grandparent papers as well), then a more thorough picture of the user community will result.

Citation Mining produced very different patterns for Fenn and Tanaka from the Bibliometrics component of the analysis. Fenn clearly stimulated the development and growth of Electrospray Ionization Mass Spectrometry, as the magnitude and timing of his citations showed.

It was unclear from the Bibliometrics that Tanaka stimulated the development and growth of soft laser desorption ionization mass spectrometry/ MALDI more than Karas and Hillenkamp. Both the early citations (from papers published in 1990-1992) and more recent citations (from papers published immediately pre-October 2002) show a more voluminous association of Karas/ Hillenkamp's early papers with soft laser desorption ionization mass spectrometry/ MALDI than Tanaka's. This issue is further exascerbated when comparing the factor matrix taxonomies of Fenn's and Tanaka's citing paper databases. There are more factors focused on applications in Fenn's citing papers, whereas there are more factors focused on mass spectrometer components in Tanaka's citing papers. A more in-depth analysis would be required to address the implications of these pattern differences, including the examination of many of the full text papers that cite Tanaka's and Karas/ Hillenkamp's works. Such an analysis was beyond the scope of the present study, but the Bibliometrics has served as an agent to flag the anomoly.

The text mining identified the major technical thrusts of both the Fenn and Tanaka citing databases. The document clustering identified both the main technical thrusts

and the number of papers devoted to each thrust. If an abbreviated text mining methodology is desired to identify major technical thrusts and approximate levels of effort devoted to each thrust, the document clustering methodology could provide a reasonable first approximation.

The main differences in the higher taxonomy levels appeared to be two-fold. First, the Tanaka citing paper applications are evenly split between bio-molecules and oligomers/ polymers, whereas the Fenn citing papers appear to focus predominately on bio-molecules. This reflects the ability of the MALDI approach to address both bio-molecules and a wide range of polymers, whereas electrospray requires soluble analytes that are readily ionizable. This restricts the classes of polymers that can be analyzed by ESI. Second, there is a MALDI component in the Fenn citing papers, but not an ESI component in the Tanaka citing papers. This reflects the practical situation that MALDI can be viewed as an alternative to ESI for bio-molecules, but ESI is much less an alternative to MALDI for polymers, for the analyte solubility reason shown above.

6. REFERENCES

1. Yamashita M, Fenn JB. Electrospray Ion-Source - Another Variation on the Free-Jet Theme. Journal of Physical Chemistry. 88 (20): 4451-4459. 1984.
2. Yamashita M, Fenn JB. Negative-Ion Production with the Electrospray Ion-Source. Journal of Physical Chemistry. 88 (20): 4671-4675. 1984.
3. Whitehouse CM, Dreyer RN, Yamashita M, Fenn JB. Electrospray Interface for Liquid Chromatographs and Mass Spectrometers. Analytical Chemistry. 57 (3): 675-679. 1985.
4. Wong SF, Meng CK, Fenn JB. Multiple Charging in Electrospray Ionization of Poly(Ethylene Glycols). Journal of Physical Chemistry. 92 (2): 546-550 Jan 28. 1988.
5. Mann M, Meng CK, Fenn JB. Interpreting Mass-Spectra Of Multiply Charged Ions. Analytical Chemistry. 61 (15): 1702-1708 Aug 1 1989.
6. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray Ionization for Mass-Spectrometry of Large Biomolecules. Science. 246 (4926): 64-71 Oct 6 1989.
7. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray Ionization-Principles and Practice. Mass Spectrometry Reviews. 9 (1): 37-70 Jan 1990.
8. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y. Protein and Polymer Analysis up to M/Zx 100000 by Laser Ionisation Time-of-Flight Mass Spectrometry. Rapid Communications In Mass Spectrometry. 2. 1988. 151-153.

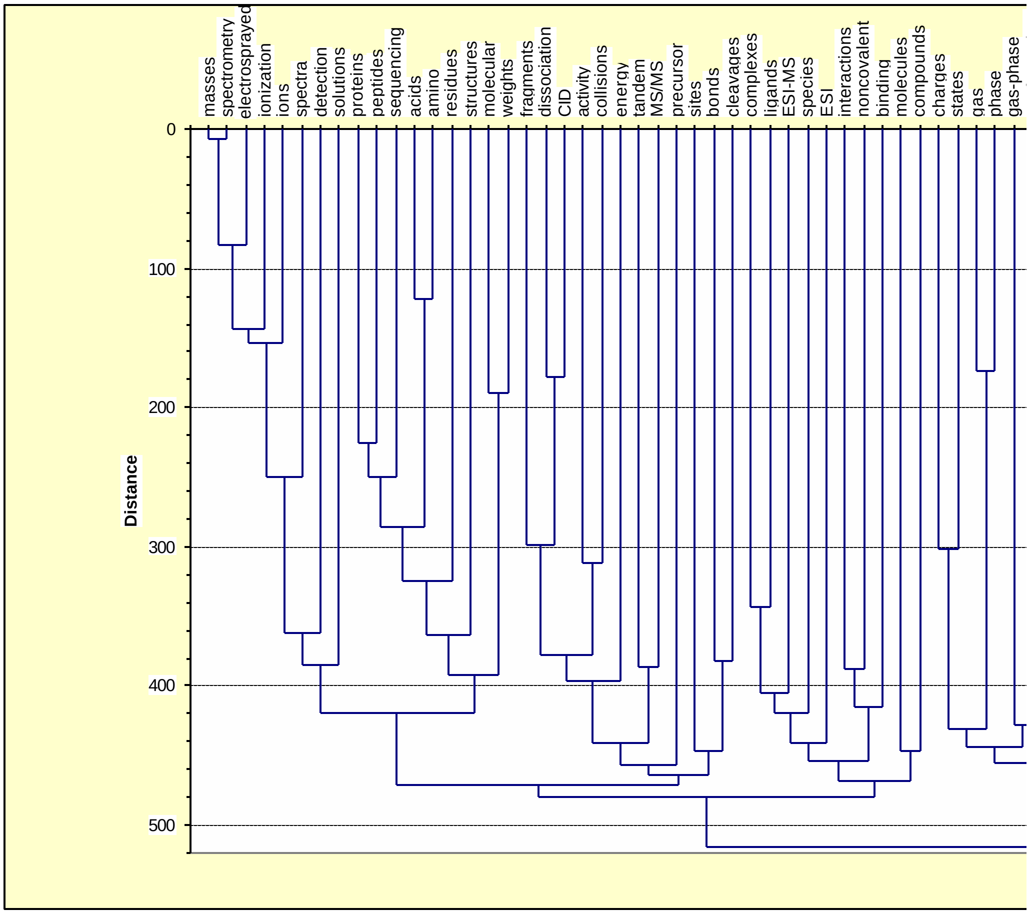
9. Tanaka, K., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T. Proceedings Second Japan-China Joint Symposium on Mass Spectrometry. Editors Matsuda, H. and Xiao-tian L. (Osaka, Japan, 15-18 September 1987). 185-188.
10. Yoshida, T., Tanaka, K., Ido, Y., Akita, S., Yoshida, Y. Mass Spectroscopy (Japan). 36. 1988. 59.
11. Beavis R. C, Chait B. T. High-Accuracy Molecular Mass Determination of Proteins Using Matrix-Assisted Laser Desorption Mass-Spectrometry. Analytical Chemistry. 62 (17): 1836-1840. Sep 1 1990.
12. Beavis, R. C., Chait, B. T. Cinnamic Acid Derivatives as Matrices for Ultraviolet Laser Desorption Mass Spectrometry of Proteins. Rapid Communications in Mass Spectrometry. 3(12). 432-435. 1989.
13. Beavis, R.C., Chait, B.T. Rapid Communications in Mass Spectrometry, 3(7), 233-237. 1989.
14. Karas M, Hillenkamp F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10000 Daltons. Analytical Chemistry. 60 (20): 2299-2301 Oct 15 1988.
15. Karas M, Bachmann D, Bahr U, Hillenkamp F. Matrix-Assisted Ultraviolet-Laser Desorption of Nonvolatile Compounds. International Journal of Mass Spectrometry and Ion Processes. 78: 53-68 Sep 24 1987.
16. Kostoff, R. N. Text Mining for Global Technology Watch. In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799. 2003.
17. Hearst, M. A. Untangling Text Data Mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.
18. Zhu DH, Porter AL. Automated Extraction and Visualization of Information for Technological Intelligence and Forecasting. Technological Forecasting and Social Change. 69 (5): 495-506 Jun 2002.
19. Losiewicz, P., Oard, D., and Kostoff, R. N. Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information Systems. 15. 99-119. 2000
20. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. Database Tomography for Information Retrieval. Journal of Information Science, 23:4, 1997.
21. Greengrass, E. Information Retrieval: An Overview. National Security Agency, TR-R52-02-96, 28 February 1997.
22. TREC (Text Retrieval Conference), Home Page, http://trec.nist.gov/.
23. Swanson, D.R. Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge. Perspect Biol Med, .30: (1), 1986.

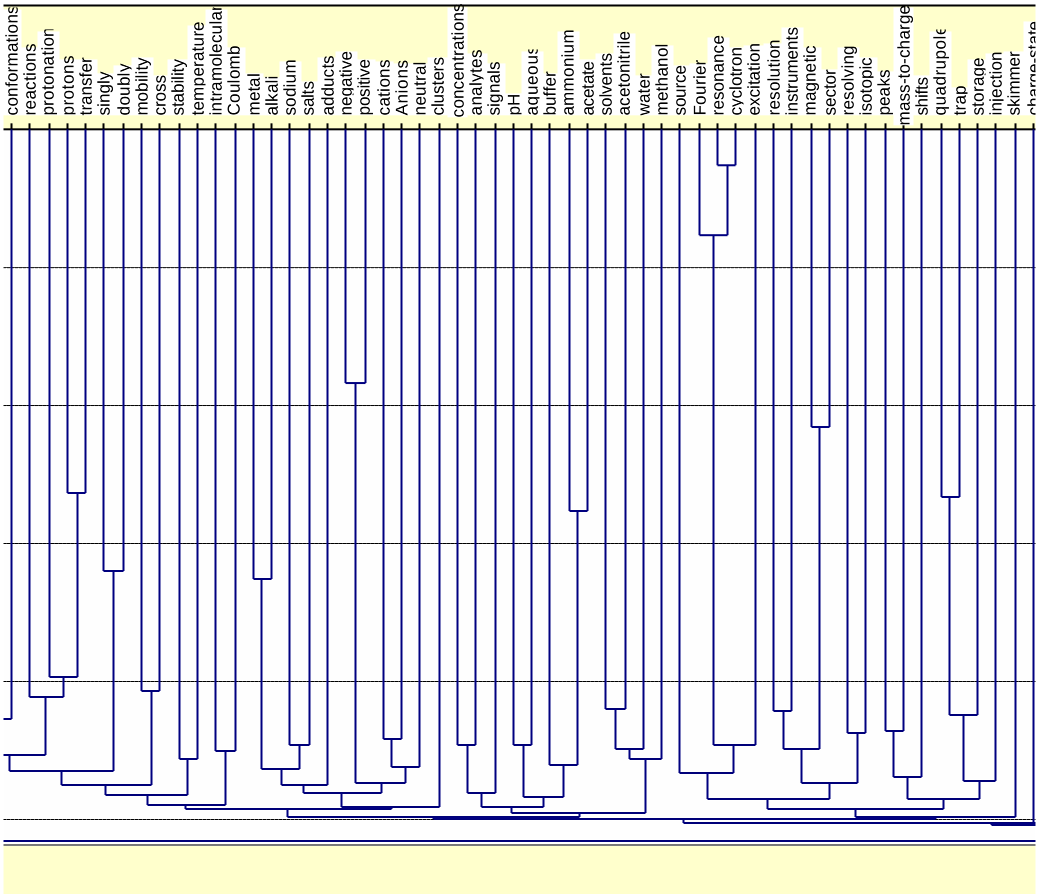
24. Swanson, D.R., Smalheiser, N.R. An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery. Artif Intell, 91 (2), 1997.
25. Kostoff, R. N. Stimulating Innovation. International Handbook of Innovation. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK. 2003.
26.Gordon MD, Dumais S. Using Latent Semantic Indexing for Literature Based Discovery. Journal of the American Society for Information Science. 49 (8): 674-685 Jun 1998.
27. Goldman, JA, Chu, WW, Parker, DS, Goldman, RM: Term domain distribution analysis: a data mining tool for text databases. Methods of Information in Medicine. 38: 96-101. 1999.
28. Kostoff, R. N. Bilateral Asymmetry Prediction. Medical Hypotheses. 61:2. 265-266. August 2003.
29. Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. Database Tomography Applied to an Aircraft Science and Technology Investment Strategy. Journal of Aircraft, 37:4. 727-730. July-August 2000.
30. Kostoff, R. N., Shlesinger, M., and Malpohl, G. Fractals Roadmaps using Bibliometrics and Database Tomography. Fractals. December 2003.
31. Viator JA, Pestorius FM . Investigating trends in acoustics research from 1970-1999. Journal of the Acoustical Society of America. 109 (5): 1779-1783 Part 1 May 2001.
32. Kostoff, R. N., Shlesinger, M., and Tshiteya, R. Nonlinear Dynamics Roadmaps using Bibliometrics and Database Tomography. International Journal of Bifurcation and Chaos. January 2004.
33. Davidse, R. J. and Van Raan, A. F. J. Out of particles: impact of CERN, DESY, and SLAC research to fields other than physics *Scientometrics* 40:2 , 171-193. 1997.
34. Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling. Journal of the American Society for Information Science and Technology. 52:13. 1148-1156. 2001.
35. Narin, F. Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity (monograph). NSF C-637. National Science Foundation. Contract NSF C-627. NTIS Accession No. PB252339/AS.
36. Garfield, E. History of Citation Indexes for Chemistry - A Brief Review. JCICS. 25(3). 170-174. 1985.

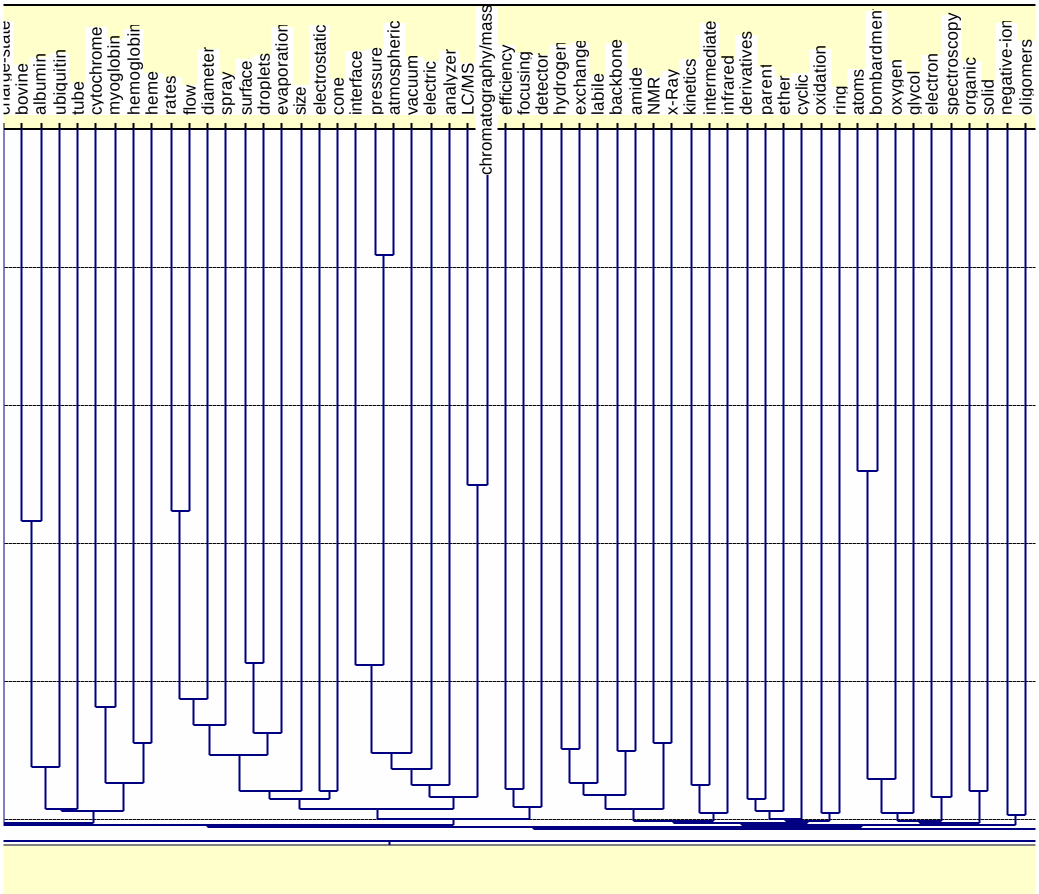
37. Schubert A, Glanzel W, Braun T. Subject Field Characteristic Citation Scores and Scales for Assessing Research Performance. Scientometrics. 12 (5-6): 267-291 Nov 1987.
38. Narin-F, Olivastro-D, Stevens-KA. Bibliometrics Theory, Practice and Problems. Evaluation Review, Vol 18, Iss 1, pp 65-76. 1994.
39. del Río, J.A. Kostoff, R.N. García, E. O., Ramírez, A. M., and Humenik, J. A. Phenomenological Approach to Profile Impact of Scientific Research. Adv. Complex Syst. 5. 19-42. 2002.
40. Kostoff, R. N., Bedford, C., Del Rio, J. A., Cortes, H. D., and Karypis, G. "Science and Technology Text Mining: Citation Mining of Macromolecular Mass Spectrometry." Journal of the American Society for Mass Spectrometry.
41. Kostoff, R. N. The Use and Misuse of Citation Analysis in Research Evaluation. Scientometrics, 43:1. September. 1998.
42. MacRoberts, M., and MacRoberts, B. Problems of Citation Analysis. Scientometrics. 36:3. July-August, 1996.
43. Smith RD, Loo JA, Edmonds CG, Barinaga CJ, Udseth HR. New Developments in Biochemical Mass-Spectrometry - Electrospray Ionization. Analytical Chemistry. 62 (9): 882-899 May 1 1990.
44. Loo JA, Edmonds CG, Smith RD. Primary Sequence Information from Intact Proteins By Electrospray Ionization Tandem Mass-Spectrometry. Science. 248 (4952): 201-204 Apr 13 1990.
45. Loo JA, Udseth HR, Smith RD. Peptide and Protein-Analysis by Electrospray Ionization Mass-Spectrometry and Capillary Electrophoresis Mass-Spectrometry. Analytical Biochemistry. 179 (2): 404-412 Jun 1989.
46. Karas M, Bachmann D, Hillenkamp F. Influence Of the Wavelength in High-Irradiance Ultraviolet-Laser Desorption Mass-Spectrometry of Organic-Molecules. Analytical Chemistry. 57 (14): 2935-2939 1985.
47. Karas M, Hillenkamp F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10000 Daltons. Analytical Chemistry. 60 (20): 2299-2301 Oct 15 1988.
48. Karas M, Bahr U, Hillenkamp F. Uv Laser Matrix Desorption Ionization Mass-Spectrometry of Proteins in the 100 000 Dalton Range. International Journal of Mass Spectrometry And Ion Processes. 92: 231-242 Sep 15 1989.
49. Jaeger, H. M., and Nagel, S. R. Physics of the Granular State. Science. 256. 20 March, p. 1523-1531. 1992.
50. Cutting, D. R., Karger, D. R, Pedersen, J. O. and Tukey, J. W. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'92). 318-329. 1992.

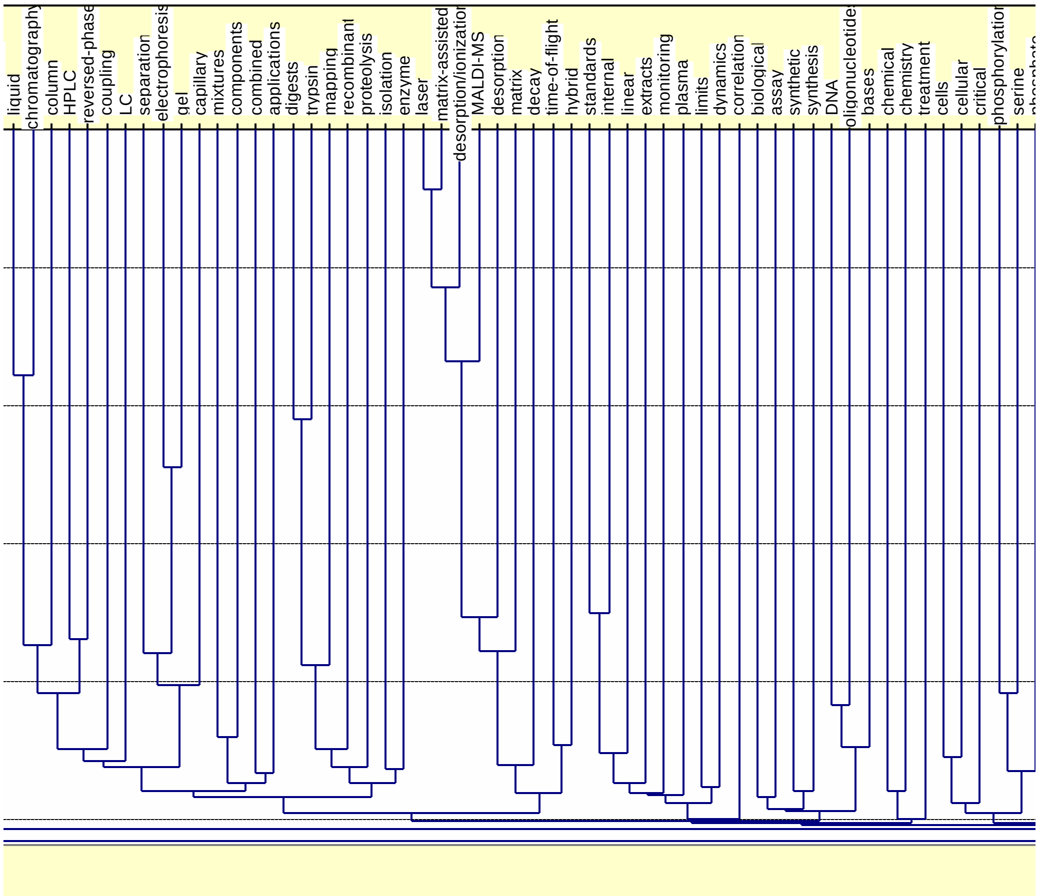
51. Guha, S., Rastogi, R. and Shim, K. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data* (SIGMOD'98). 73-84. 1998.
52. Hearst, M. A. The use of categories and clusters in information access interfaces. In T. Strzalkowski (ed.), Natural Language Information Retrieval. Kluwer Academic Publishers. 1998.
53. Karypis, G.; Han, E.-H.; and Kumar, V. Chameleon: A hierarchical clustering algorithm using dynamic modeling. IEEE Computer: Special Issue on Data Analysis and Mining 32(8). 68--75. 1999.
54. Prechelt, L., Malpohl, G., Philippsen, M. Finding plagiarisms among a set of programs with JPlag. Journal of Universal Computer Science 8(11). 1016-1038. 2002.
55. Rasmussen, E. *Clustering Algorithms*. In W. B. Frakes and R. Baeza-Yates (eds.). 1992]. Information Retrieval Data Structures and Algorithms, Prentice Hall, N. J.
56. Steinbach, M.; Karypis, G.; and Kumar, V. A comparison of document clustering techniques. Technical Report #00--034. Department of Computer Science and Engineering. University of Minnesota. 2000.
57. Willet, P. Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*. 24:577-597. 1988.
58. Wise, M. J. String similarity via greedy string tiling and running Karb-Rabin matching. ftp://ftp.cs.su.oz.au/michaelw/doc/RKR_GST.ps, Dept. of CS, University of Sidney. 1992.
59. Zamir, O. and Etzioni, O. Web document clustering: A feasibility demonstration. In: *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'98). 46-54. 1998.
60. Karypis, G. (2002). CLUTO—A clustering toolkit. http://www.cs.umn.edu/˜cluto.
61. Y. Zhao and G. Karypis (2003) Criterion functions for document clustering: Experiments and analysis. Machine Learning, in press

FIGURE 2 – DENDOGRAM – FENN CITING PAPERS


masses
spectrometry
electrosprayed
ionization
ions
spectra
detection
solutions
proteins
peptides
sequencing
acids
amino
residues
structures
molecular
weights
fragments
dissociation
CID
activity
collisions
energy
tandem
MS/MS
precursor
sites
bonds
cleavages
complexes
ligands
ESI-MS
species
ESI
interactions
noncovalent
binding
molecules
compounds
charges
states
gas
phase
gas-phase
Distance
0
100
200
300
400
500


conformations
reactions
protonation
protons
transfer
singly
doubly
mobility
cross
stability
temperature
Coulomb
metal
alkali
sodium
salts
adducts
negative
positive
cations
Anions
neutral
clusters
concentrations
analytes
signals
pH
buffer
ammonium
acetate
solvents
acetonitrile
water
methanol
source
Fourier
resonance
cyclotron
excitation
resolution
instruments
magnetic
sector
resolving
isotopic
peaks
mass-to-charge
shifts
trap
storage
injection
skimmer


bovine
albumin
ubiquitin
tube
cytochrome
myoglobin
hemoglobin
heme
rates
flow
diameter
spray
surface
droplets
evaporation
size
electrostatic
cone
interface
pressure
atmospheric
vacuum
electric
analyzer
LC/MS
chromatography/mass
efficiency
focusing
detector
hydrogen
exchange
labile
backbone
amide
NMR
x-Ray
kinetics
intermediate
infrared
derivatives
parent
ether
cyclic
oxidation
ring
atoms
bombardmen
oxygen
glycol
electron
spectroscopy
organic
solid
negative-ion
oligomers


liquid
chromatography
column
HPLC
reversed-phase
coupling
LC
separation
electrophoresis
gel
capillary
mixtures
components
combined
applications
digests
trypsin
mapping
recombinant
proteolysis
isolation
enzyme
laser
matrix-assisted
desorption/ionization
MALDI-MS
desorption
matrix
decay
time-of-flight
hybrid
standards
internal
linear
extracts
monitoring
plasma
limits
dynamics
correlation
biological
assay
synthetic
synthesis
DNA
oligonucleotides
bases
chemical
chemistry
treatment
cells
cellular
critical
phosphorylation
serine

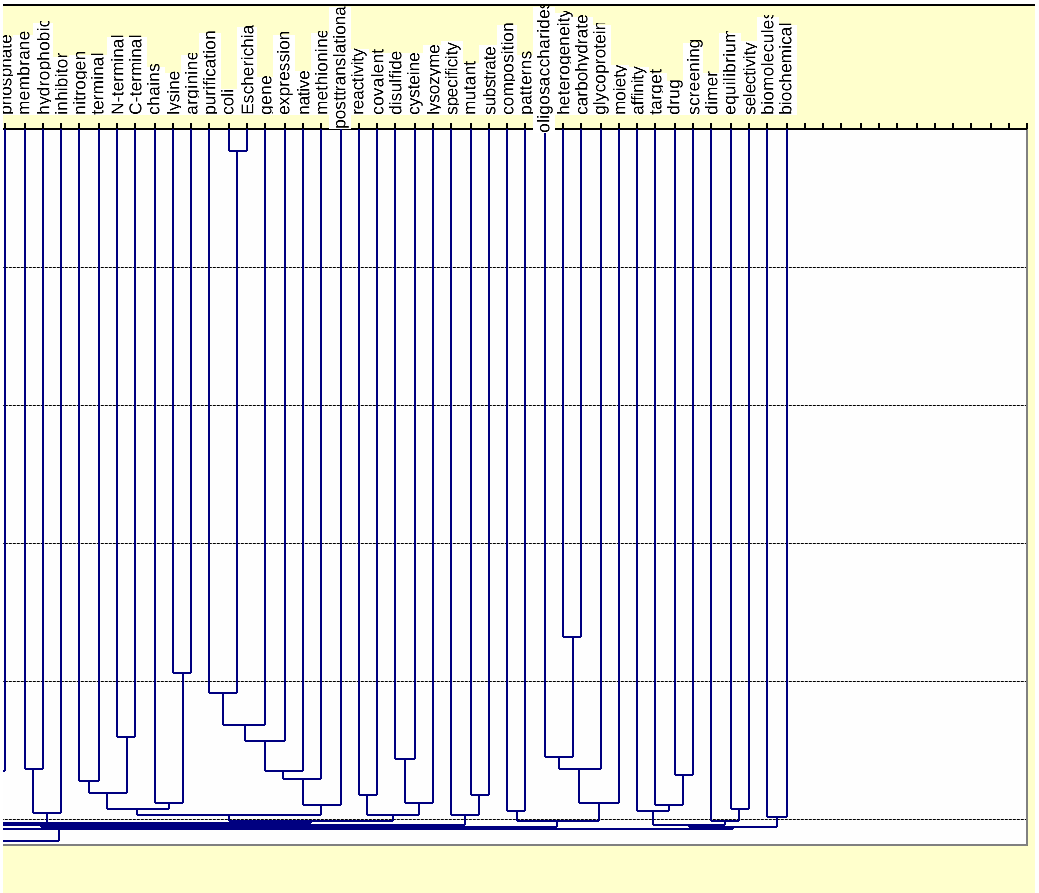
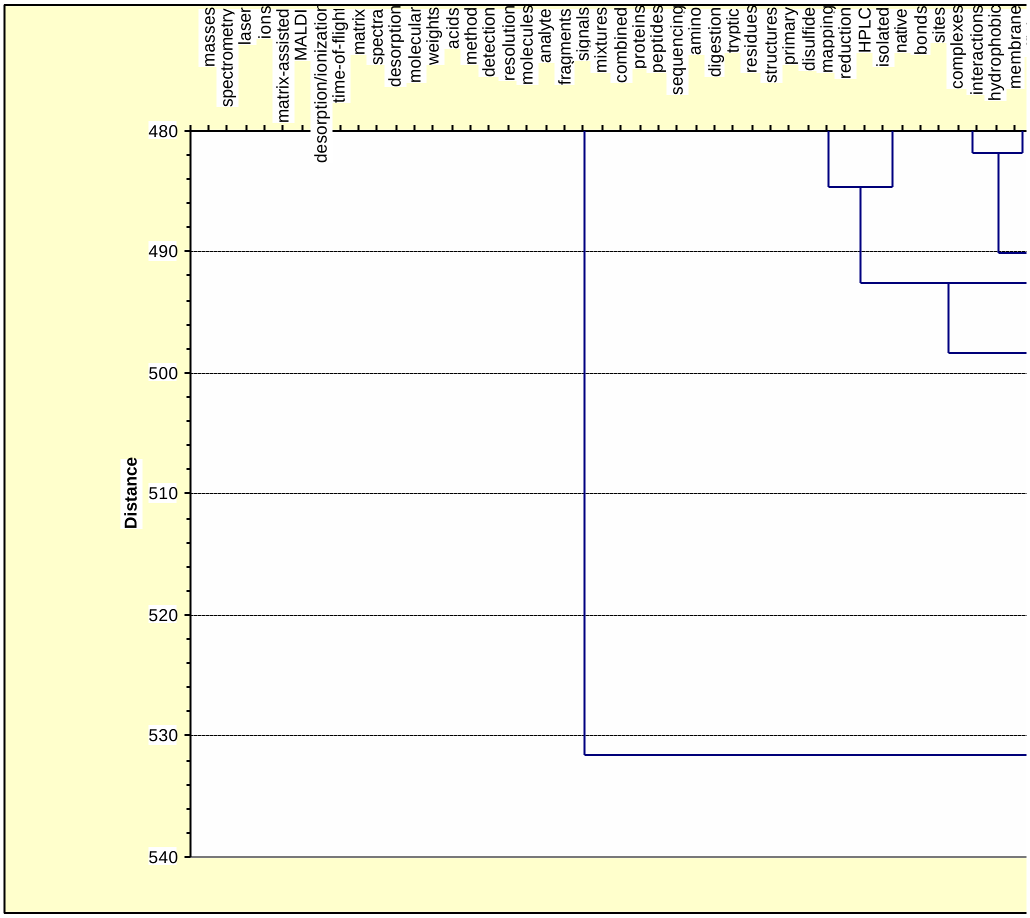
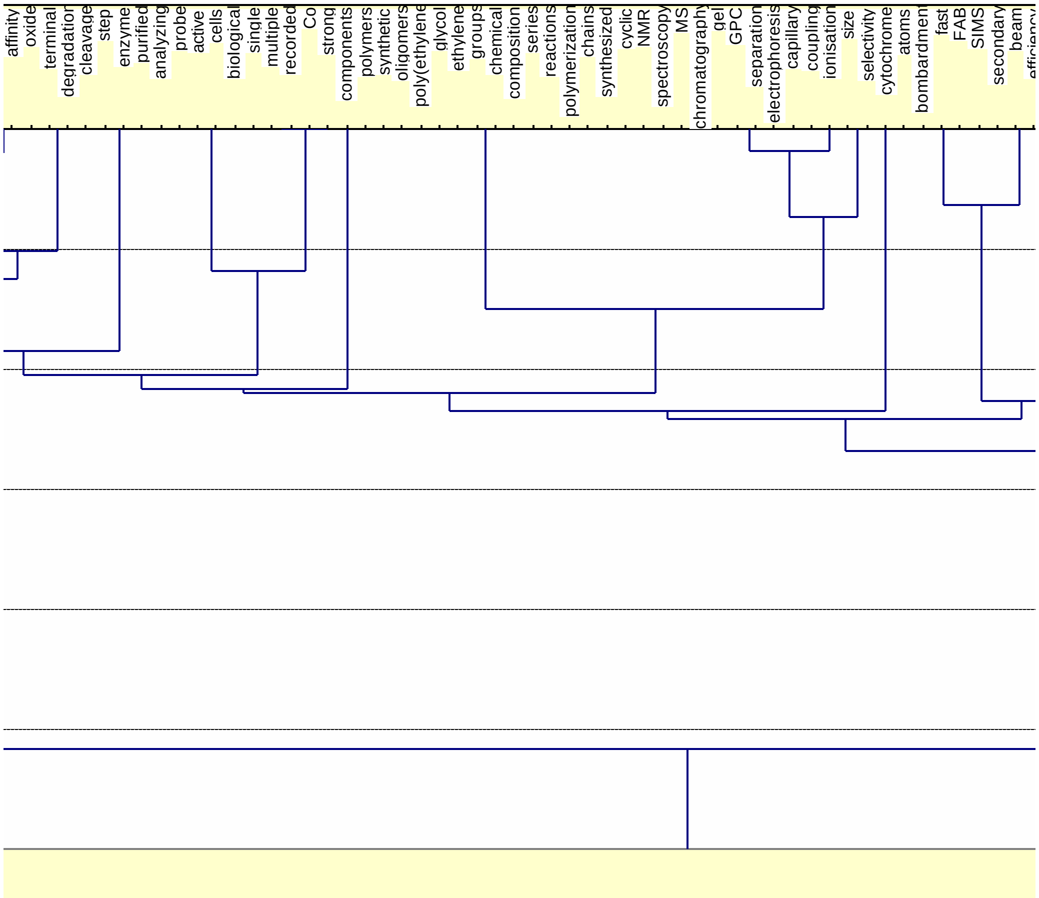
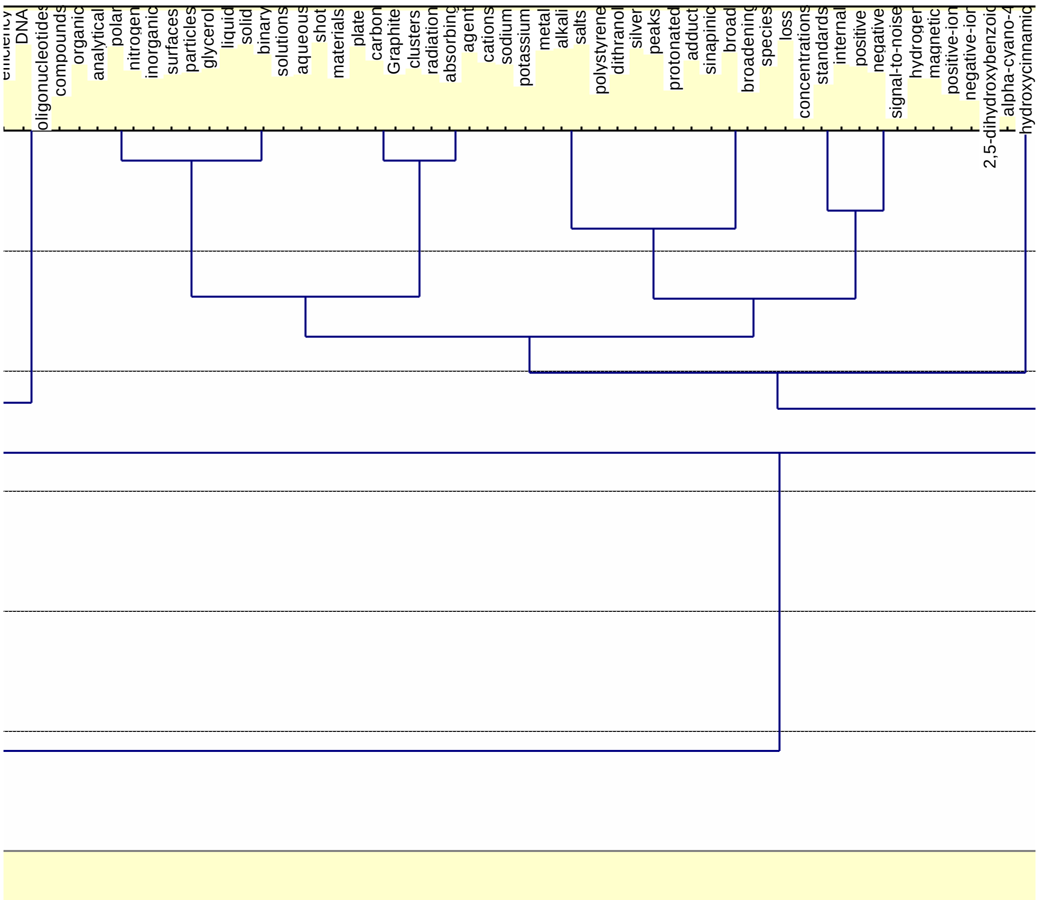

phosphate
membrane
hydrophobic
inhibitor
nitrogen
terminal
N-terminal
C-terminal
chains
lysine
arginine
purification
coli
Escherichia
gene
expression
native
methionine
posttranslational
reactivity
covalent
disulfide
cysteine
lysozyme
specificity
mutant
substrate
composition
patterns
oligosaccharides
heterogeneity
carbohydrate
glycoprotein
moiety
affinity
target
drug
screening
dimer
equilibrium
selectivity
biomolecules
biochemical

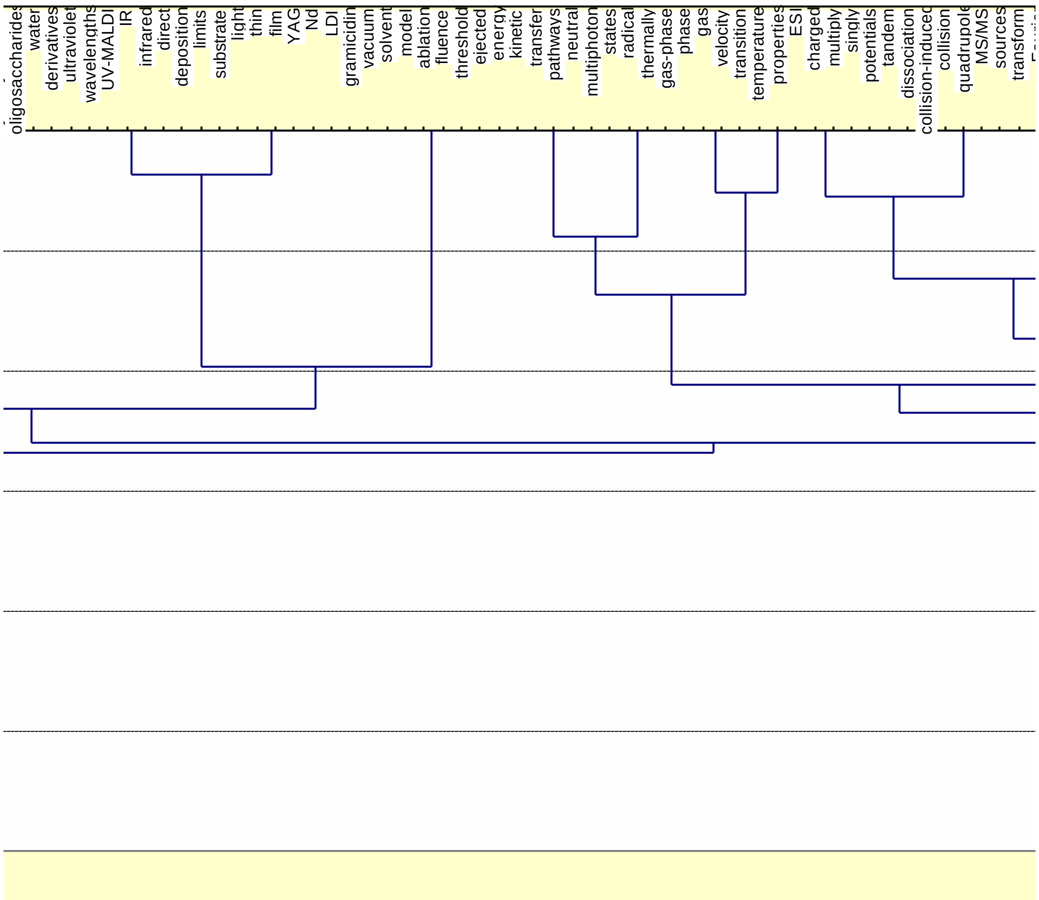
FIGURE 3 – DENDOGRAM - TANAKA CITING PAPERS

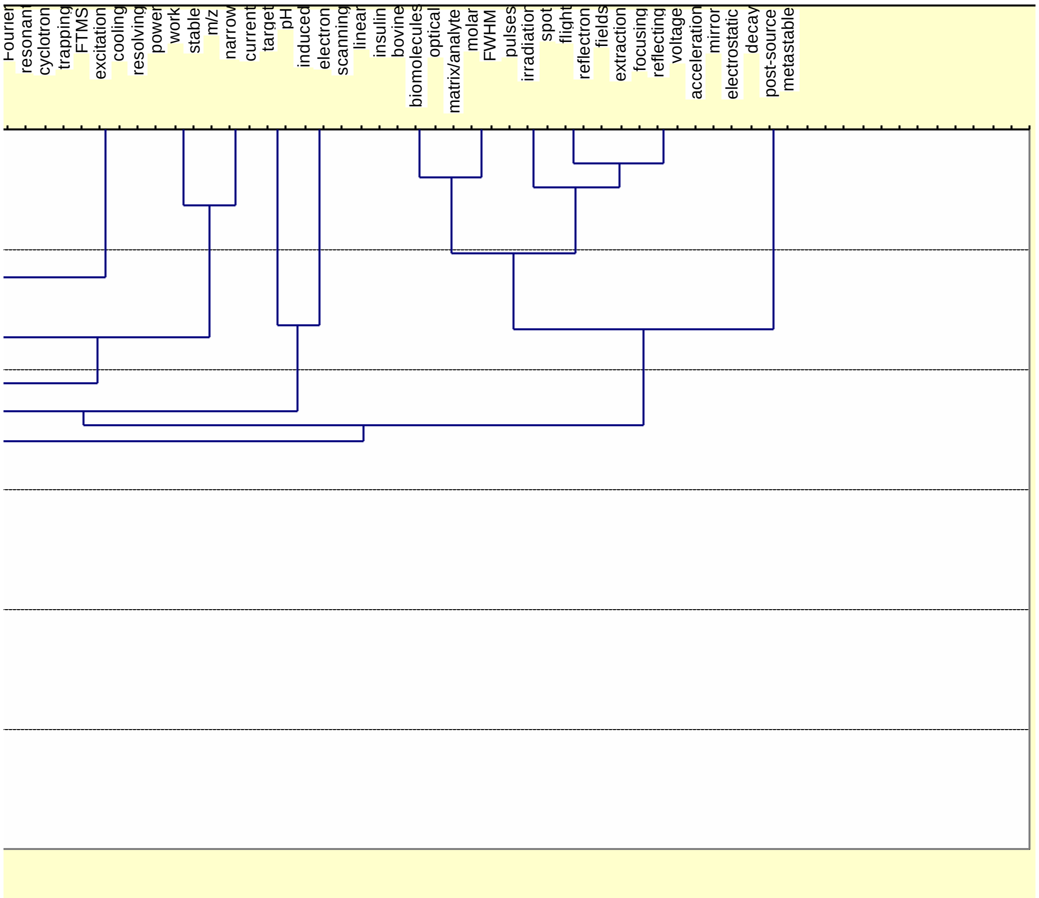



affinity
oxide
terminal
degradatio
cleavage
step
enzyme
purified
analyzing
probe
active
cells
biological
single
multiple
recorded
Co
strong
components
polymers
synthetic
oligomers
poly(ethylene
glycol
ethylene
groups
chemical
composition
series
reactions
polymerization
chains
synthesized
cyclic
NMR
spectroscopy
MS
chromatography
gel
GPC
separation
electrophoresis
capillary
coupling
ionisation
size
selectivity
cytochrome
atoms
bombardmen
fast
FAB
SIMS
secondary
beam



DNA
oligonucleotides
compounds
organic
analytical
polar
nitrogen
inorganic
surfaces
particles
glycerol
liquid
solid
binary
solutions
aqueous
shot
materials
plate
carbon
Graphite
clusters
radiation
absorbing
agent
cations
sodium
potassium
metal
alkali
salts
polystyrene
dithranol
silver
peaks
protonated
adduct
sinapinic
broad
broadening
species
loss
concentrations
standards
internal
positive
negative
signal-to-noise
hydrogen
magnetic
positive-ion
negative-ion
2,5-dihydroxybenzoic
alpha-cyano-4
hydroxycinnamic


oligosaccharides
water
derivatives
ultraviolet
wavelengths
UV-MALDI
IR
infrared
direct
deposition
limits
substrate
light
thin
film
YAG
Nd
LDI
gramicidin
vacuum
solvent
model
ablation
fluence
threshold
ejected
energy
kinetic
transfer
pathways
neutral
multiphoton
states
radical
thermally
gas-phase
phase
gas
velocity
transition
temperature
properties
ESI
charged
multiply
singly
potentials
tandem
dissociation
collision-induced
collision
quadrupole
MS/MS
sources
transform



Fourier
resonant
cyclotron
trapping
FTMS
excitation
cooling
resolving
power
work
stable
m/z
narrow
current
target
pH
induced
electron
scanning
linear
insulin
bovine
biomolecules
optical
matrix/analyte
molar
FWHM
pulses
irradiation
spot
flight
reflectron
fields
extraction
focusing
reflecting
voltage
acceleration
mirror
electrostatic
decay
post-source
metastable